Helsinki Studies in Education, number 87

Ninja Hienonen

Does class placement matter?
Students with special educational needs in regular and special classes

To be presented, with the permission of the Faculty of Educational Sciences of the University of Helsinki, for public discussion in Unioninkadun juhlasali 303, Unioninkatu 33, on Friday September 25th, 2020 at 12 noon.

Helsinki 2020

Pre-examiners
Professor Emeritus Jan Van Damme, KU Leuven
Professor Rauno Parrila, Macquarie University

Custos
Professor Markku Jahnukainen, University of Helsinki

Supervisors
Professor Markku Jahnukainen, University of Helsinki
Professor Risto Hotulainen, University of Helsinki
Associate Professor Mari-Pauliina Vainikainen, Tampere University

Opponent
Professor Mikko Aro, University of Jyväskylä

Cover
Lamminsalon koulu 1936. Photographer unkown.

The Faculty of Educational Sciences uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

Yliopistopaino Unigrafia, Helsinki

ISBN 978-951-51-6391-2 (nid.)
ISBN 978-951-51-6392-9 (PDF)

University of Helsinki, Faculty of Educational Sciences
Helsinki Studies in Education, number 87

Ninja Hienonen

Does class placement matter?
Students with special educational needs in regular and special classes

Abstract

Faced with a diverse student population, schools assign student into classes of different size and composition. These placements can have consequences on learning and teaching and they are often referred to as compositional effects. Consequently, in this study I treated students in clusters (classes in schools) with a hypothesis that students can expect to have different levels of performance depending on the class they are assigned to. The main focus was on students with special educational needs (SEN), and on the question of how they are affected by the class placement.

The aim of the study was to discern the class-level effects, specifically, class size and the proportion of students with SEN in regular classes, and to explore the effect of the placement by comparing regular and special classes. Data were drawn from two longitudinal large-scale learning to learn assessment studies representing both primary (N = 896) and lower secondary education (N = 5368). Data were analyzed with multilevel regression models. Furthermore, quasi-experimental design was created using propensity score matching.

The results of this study confirmed that on average, students with SEN were placed in smaller classes, however, the size of a class as such had no effect on their performance in cognitive tasks. Furthermore, the average performance level in regular classes with students with SEN was lower than in classes without students with SEN, and the effect remained significant even after the initial differences were controlled for. Students with SEN seemed to benefit from the other students with SEN placed in the same classroom. In addition, the results suggested a tendency to create more homogeneous classrooms as less-achieving students without SEN were placed together with students with SEN.

When the differences among students with SEN placed in two distinct educational settings, special and regular classes were explored, no differences in any cognitive tasks were detected. However, students in special classes received higher grades in some core subjects, and that calls for more research on grading practices in different classroom contexts. The findings also revealed differences in learning motivation across the two settings.

The purposeful sorting of all students, not only students with SEN into classrooms was confirmed with this study. The results also implied a hidden tracking system within schools. It is evident that assigning students into

classrooms is far from neutral act, and that there can be some unintentional consequences. The criteria that are used in assigning students to classes should be discussed in an explicit manner and the schools and administrators should be aware of the possible consequences of different placement decisions.

Keywords: students with special educational needs, class composition, regular class, special class

Helsingin yliopisto, Kasvatustieteellinen tiedekunta
Kasvatustieteellisiä tutkimuksia, numero 87

Ninja Hienonen

Onko oppilaan luokalla merkitystä?

Tukea saavat oppilaat yleisopetuksen ja erityisopetuksen luokissa

Tiivistelmä

Hallitakseen heterogeenistä oppilasjoukkoa koulut jakavat oppilaita eri kokoisiin ja erilaisin perustein ryhmiteltyihin luokkiin. Näillä opetuspaikkaratkaisuilla on yhteys oppimiseen ja opettamiseen. Tätä yhteyttä voidaan kutsua luokkakomposition vaikutukseksi. Tässä tutkimuksessa oppilaita käsitellään koululuokkien mukaisissa ryhmissä ja keskitytään pääosin tehostettua tai erityistä tukea saavien oppilaiden toteutuspaikkaratkaisuihin. Tutkimus testaa hypoteesia, jonka mukaan oppilaan oppimistulosten voidaan olettaa osin riippuvan siitä, millaisessa luokassa oppilas opiskelee. Asiaa tarkastellaan luokkakoon ja luokanmuodostuksen näkökulmasta. Lisäksi tutkitaan erityistä tukea saavia oppilaita suhteessa siihen, ovatko he yleisopetuksen luokilla vai erityisluokalla.

Aineistona käytetään pitkittäistutkimuksena kerättyjä oppimaan oppimisen oppilasaineistoja perusopetuksen ala- ja yläluokilta (N = 896 ja N = 5368). Analyysimenetelmänä käytetään pääosin monitasomallinnusta. Lisäksi luodaan kvasikokeellinen asetelma hyödyntäen parinkaltaistustekniikkaa.

Tutkimuksen tulokset osoittivat, että tukea saavat oppilaat opiskelevat keskimäärin pienimmissä luokissa. Luokkakoolla ei kuitenkaan ollut suoraa yhteyttä oppilaiden suoriutumiseen osaamistehtävissä. Niissä yleisopetuksen luokissa, joissa opiskeli tukea saavia oppilaita, luokkatasoinen suoriutuminen osaamistehtävissä oli keskimäärin heikompaa kuin muilla luokilla ja tämä yhteys säilyi, vaikka oppilaiden väliset lähtökohtaiset osaamiserot huomioitiin. Tukea saavat oppilaat näyttivät hyötyvän siitä, että samassa luokassa oli muitakin tukea saavia oppilaita. Lisäksi tulokset antoivat viitteitä siitä, että koulut jakavat oppilaita luokkiin pyrkien luomaan mahdollisimman yhtenäisiä luokkia. Oppilaat, jotka opiskelivat yleisopetuksen luokissa yhdessä tukea saavien oppilaiden kanssa, erosivat muiden oppilaiden luokista heikommalla lähtötasollaan.

Erityistä tukea saavat oppilaat eivät eronneet osaamistehtävissä sen perusteella, opiskelivatko he pääsääntöisesti yleisopetuksen luokalla vai erityisluokissa. Erityisluokilla oppilailla oli kuitenkin korkeampia arvosanoja. Tämä tulos osoittaa selkeän lisätutkimuksen tarpeen arviointikäytänteisiin liittyen. Lisäksi oppilaiden oppimismotivaatiossa havaittiin eroja.

Tämä työn perusteella voidaan todeta, että tukea saavien oppilaiden toteutuspaikkaratkaisut koskettavat kaikkia oppilaita. Tulokset antoivat myös viitteitä piilevistä tasoryhmien kaltaisista ryhmittelyistä. Selvää on, että kouluilla

on oltava keinoja hallita moninaista oppilasjoukkoa luokanmuodostuksen keinoin. On kuitenkin tiedostettava, että näillä opetuspaikkaratkaisuilla voi olla ennalta-arvaamattomia seurauksia. Nämä seuraukset on hyvä tiedostaa kouluissa ja keskustella niistä avoimesti.

Avainsanat: tukea saavat oppilaat, luokanmuodostus, yleisopetuksen luokka, erityisluokka

Dedicated to my godson Sebastian, a future student

Acknowledgements

Writing a dissertation can sometimes be lonely; yet thankfully, it is not done alone. I feel privileged to have so many people around me who have made this project possible. This section of this book is devoted to thanking you all.

First, I would like to express my gratitude to my supervisors. Professor Markku Jahnukainen, thank you for always having time to respond to my numerous questions along this journey. Thank you also for the autonomy when conducting my research and for encouraging me to let go of the articles when I got stuck in the details. I also want to thank you for encouraging me to excel myself. Professor Risto Hotulainen, I wish to thank you for your support and guidance, and also, for every beer that we've had during this process. Associate professor Mari-Pauliina Vainikainen, thank you for believing in me as an emergent researcher. I am also grateful for guiding me with research methods and showing me how much fun it can be.

I wish to thank the pre-examiners of this dissertation, Professor Emeritus Jan Van Damme and Professor Rauno Parrila for your encouraging words, valuable comments and also, critical feedback on the manuscript. I would also like to thank Professor Mikko Aro for agreeing on being my opponent.

I find myself fortunate for having been able to work in an environment in which I can learn something new every day. I owe my next gratitude to Sirkku Kupiainen. Thank you for sharing your expertise with me, for all the work we have done together and also, for all the late-night discussion at the university over the years. I am thankful for Professor Emeritus Jarkko Hautamäki. Thank you for making me part of your research group years ago as a young master's student. Your abstract way of thinking and attitude towards challenges has taught me much.

My critical thinking partner, Meri Lintuvuori, I am thankful that I have been able to think aloud with you. I am thankful for all your support, understanding, listening and your friendship that goes far beyond the walls of the university. I want to thank Mikko Asikainen for always helping me with the tricky matrixes, syntaxes, and even with embarrassing ICT-problems, for teaching me how to catch a fish instead of giving me one, and for all the morning porridges at Olivia. I am also thankful for our downstairs group, Esko Lindgren and Jukka Määttänen, in addition to Meri and Mikko, thank you for a supportive working climate which makes you laugh, loud and often.

Recent years, I have divided my time between the universities Helsinki and Tampere. There is an amazing bunch of other colleagues at both universities I work at: Raisa Ahtiainen, Sanna Oinas, Satu Koivuhovi, Irene Rämä, Helena Thuneberg, Milja Saarnio, Päivi Nilivaara, Laura Nyman, Marja-Liisa Kieksi, Natalija Gustavson, I am thankful for having been able to work with you. This project would have not been possible without a wider research community, the

Centre for Educational Assessment, and the people around it, you all deserve my thanks.

Along the way, I have been fortunate to meet people that have influenced my work and thinking. I would like to express my gratitude to Jorma Kuusela whose human and thorough thinking has profoundly affected my statistical understanding. I wish to express my gratitude to Jukka Marjanen, for sharing your knowledge on research methods, your patience and for making statistics understandable. I would like to thank Professor Hannu Salmi for your energy, new ideas and support. I want to express my gratitude to Professor Paul Ilsey for guiding and encouraging me and making me feel like anything is possible. I am also thankful for Professor Joseph Gagnon for providing me with valuable comments on the article manuscript. I want to thank Cristiana Mergianian for proof-reading my texts.

I wish to thank to the unknown peer reviewers and to the editors for their valuable comments that helped to shape and improve the original articles. This study owes to all the students, teachers and principals for taking part in the original data collections—without their time and effort this research project would have not been possible.

It seems that there is a life outside academia and this life has been the necessary counterforce during this process. I would like to give my sincerest thanks to my best friend Annukka Alivaara. Thank you for sharing lunch with me on the first day at the university. My dear friends Tiina Mäkelä, Eerika Ainamo, Iina Väisänen, my cousin Aleksi Levelä, my in-laws Leena, Emil, Irina and Erkko. Thank you for sharing all the ups and downs with me.

I would like to express my warmest gratitude to both my families, birth and in-law. My parents Maritta and Jukka, you managed to bring up a daughter with a desire for knowledge and lifelong learning. Thank you for always supporting my choices in life. My parents-in-law, Pirjo and Erkki, thank you for your encouragement and support. This dissertation owes much to my sister Nanne. Of all people, you have challenged my thinking most over the years. Thank you for your love and support, and for the calls that can last for hours. I would like to thank Jyri for always making your home a welcoming place for me and for all the cooking over the years. My godson Sebastian, thank you for bringing a sincere joy to my life and for reminding auntie for what really matters in life.

Finally, I wish to express all of my gratitude to Erno, who became my husband along this journey. I am grateful for your devotion, support and unquestioning believe in me. Thank you for making me step out from the computer and making me do squats. Thank you for sharing this project and life with me. I am truly blessed to have you by my side.

In Helsinki, August 2020
Ninja Hienonen

Contents

List of original articles

This thesis is based on the following three articles, which are referred to in the text by their Roman numerals (Studies I–III):

I. Vainikainen, M-P., Hienonen, N. & Hotulainen, R. (2017). Class size as a means of three-tiered support in Finnish primary schools. *Learning and Individual Differences*, *56,* 96–104.

II. Hienonen, N., Lintuvuori, M., Jahnukainen, M., Hotulainen, R. & Vainikainen, M.-P. (2018). The effect of class composition on cross-curricular competencies—Students with special educational needs in regular classes in lower secondary education. *Learning and Instruction, 58*, 80–87.

III. Hienonen, N., Hotulainen, R., & Jahnukainen, M. (2020). Outcomes of students with special educational need sin regular versus special classes: A quasi-experimental study. *Scandinavian Journal of Educational Research*.

Author's contribution

In the first article, Ninja Hienonen shared the first authorship with Mari-Pauliina Vainikainen. Hienonen served as the corresponding author. She also did the data analysis, and took the main responsibility of writing, submitting and revising the article. In the second and third article, Hienonen served as a first author. She planned the research design did the data analyses, and took the main responsibility of writing, submitting and revising the article.

1 Introduction

> "Intrinsically motivated research projects - - - start out of curiosity. One begins to wonder about a certain phenomenon, continues by questioning, then tries to find answers, and finally wants to give some new explanations for the phenomenon" (Thuneberg, 2007).

The above quotation captures the essence of how this research project started when five years ago, I wrote the first draft of a research plan I wanted to follow. I had already started a study of class size in Finland with my colleague (Kupiainen & Hienonen, 2016), and I discovered some highly important questions that were still unanswered in the Finnish context. The impetus for this research was clear, I saw a great demand for research-based knowledge of how students with special educational needs (SEN) are affected by the class placement, class size and class composition.

The questions of the class placement and class composition effects are not new; in fact, they are both timely and perpetual. T.S. Eliot, a poet and a social critic, wrote as long ago as 1933 in his essays on modern education (p. 509):

> "Anyone who has taught children even for a few weeks knows that the size of a class can make an immense difference to the amount you can teach. Fifteen is an ideal number; twenty is the maximum; with thirty much less can be done; with more than thirty most teachers' first concern is simply keep order, and the clever children creep at the pace of the backward."

Furthermore, in 1937, H. H. Postel contemplated the question of the placement of students with SEN in *Exceptional Children*. He concluded that homogenous groupings in segregated settings are the most adaptable and less stigmatizing solutions for students who struggle with learning, as they need an elastic type of organization to meet their emotional, physical, and educational needs. However, he also admits that "- - that some teachers of the single special class can surmount the difficulties presented by such a group in a regular school" (p. 19).

The world has changed since then, but the main questions remain. Never fully resolved, it seems they must be revisited by every generation (Kauffman, Nelson, Simpson, & Ward, 2017). I quickly realized I had set quite a challenging task for myself. There was a vast array of research with contradictory findings, especially on class size, and at the same time, scholarly work in the Finnish context was almost non-existent. The study by Kupiainen and Hienonen (2016) was the first to address this question in-depth. The present study continues this direction by

being the first study in Finland in which class size, the proportion of students with SEN, learning outcomes and learning motivation have all been studied.

It is said that the more you know about a certain topic, the more you learn what you do not know. This is exactly what has happened since I started to solve the puzzle of class placement and class composition in the context of Finnish education. The more I read the earlier research, the more factors, features and dimensions of the phenomenon were revealed. Thus, conducting research means seeking a constant balance and trade-offs between decisions on what to include and what to exclude. Even though the research findings on class size have indeed been inconsistent, there seemed to be a strong and persistent belief in the power of class size and class size reduction (Hattie, 2005). For most people, class size is intuitively linked to academic outcomes (Schanzenbach, 2010). In fact, class size reduction is one of the most often proposed solutions to educational challenges across the world. It is also a topic that regularly evokes political debate, heated discussions among teachers and parents, policymakers, and statements by the Trade Union of Education in Finland. It seems that everyone involved in education and schooling has an opinion on the matter. The main arguments usually are that class sizes are too large, and that both the teaching and learning can suffer (e.g., Blatchford & Russel, 2019).

There have also been debates about the increase in more challenging student populations in regular classes, including students with SEN. Consequently, a related issue is the placement of students with SEN. It is at least as heated and polarized, and somewhat emotionally driven. Being placed in a regular class with same-aged peers is seen as every student's indisputable right (e.g., European Union, 2018; UNESCO, 2017; United Nations, 2006). However, at the same time, the placement of students with SEN in regular classes among peers can be seen, at the worst, as a cost-cutting effort by the education provider (Honkasilta, Ahtiainen, Hienonen, & Jahnukainen, 2019). The aim of this study was to integrate these two topics. Furthermore, the intention was to go beyond the different views and opinions by using large-scale data and sophisticated statistical methods to add to the understanding of placement effects. Despite the urgency and importance of the question on the placement of students with SEN, it has been the subject of little objective investigation and thus, many of the placement practices do not rest clearly on research-based facts (Kauffman, Nelson et al., 2017). The topic of this dissertation is not restricted to Finland, but it is a global issue as well. Hence, the three articles that make up this dissertation were published in international education journals with the intention of adding to the continuing international discussion and to contribute to the Finnish perspective.

This research project was partly initiated by interest in class size and its effects, especially on students with SEN, as previous findings in the class size literature indicated that the lower performing students if any, could benefit from a smaller number of class mates. Yet, not even this finding is indisputable. When it turned

out that class size lacks the power to explain between-class differences in terms of student performance, motivation, or classroom climate (Kupiainen & Hienonen, 2016), my research interest shifted from the size of a class to the composition of a class, focusing on students with SEN. Study I tests the assumption that students with SEN would benefit from smaller classes. The focus of Study II is on the proportion of students with SEN in regular classes, its relation to the class-level performance, and to its relation to students with SEN and to students without SEN. Studies I and II are centered around the regular classes but taking into account the students without SEN, whereas Study III focuses on students with SEN placed either in regular or special classes. In Study III, the class size also plays a part as the class size maximum in special classes is regulated by law.

Characteristics of the Finnish education from an international perspective have been the 21st century success in international comparisons in terms of school attainment assessments (Mullis, Martin, Foy, & Hooper, 2017; OECD, 2012; OECD, 2016), the decentralized education system (Varjo & Kalalahti, 2019), and strong teacher autonomy (Niemi, 2015). Furthermore, the extensive Finnish special education system has been seen as a distinctive feature in basic education and, as an explanation of the fairly unique system in which the differences between school and student performance have been small, and where the lowest performing Finnish students have outperformed their comparison groups in other Organisation for Economic Co-operation and Development (OECD) countries (Kivirauma & Ruoho, 2007; OECD, 2016a) until recently. The declining attainment results, increasing differences between the lowest and highest-performing students, and the growing proportion of lowest-performing students have aroused concerns. The extent to which these worrying signs could be explained by the challenges in the special education system has been discussed (Vettenranta et al., 2016). In addition, while the differences between Finnish schools have been extremely small, the differences between classes have been high when compared to other OECD countries (Yang Hansen, Rósen, & Gustafsson, 2014). Furthermore, Finland stands out in comparison to other OECD counterparts as its primary education teachers most often report having higher proportions (more than 60%) of low academic achievers in their class, and higher rates of students with SEN in classrooms (OECD, 2014, p. 44; also, OECD, 2019b). Clearly, attention must be paid to the class level in Finnish schools and to the placement of students with SEN.

Classes in schools are not free-standing units, since their formation represents the result of administrative decisions by which grades are subdivided into smaller units based on different decisions. Thus, grouping students is not a neutral act, rather it is a potential arm of educational policy (Harker & Tymms, 2004). It is clear that sorting of students into classrooms is one way to manage student diversity in schools and to respond to initial student differences (e.g., Harker &

Tymms, 2004). Essentially, by assigning students into sub-groups that are more homogeneous than the population as a whole, schools run like many other complex organizations (Dreeben & Barr, 1988). Furthermore, it is commonly believed that all organizations can accomplish their goals more efficiently when they allocate separate tasks to specialized sub-units (Gamoran, Nystrand, Berends, & Lepore, 1995). In particular, students with SEN face a variety of placement options as a placement can be anything from a full-time placement in a regular class to a full-time placement in a special class in a special school; therefore, the continuum of special education contexts is broader than general education. The placement of students with SEN in different types of class can be seen as a kind of ability grouping (Myklebust, 2007) and thus, the effects of the placement must be put under scrutiny. The high between-class differences in Finnish schools have already been acknowledged (e.g., Kupiainen, 2019; Thuneberg, Hautamäki, & Hotulainen, 2015; Yang Hansen et al., 2014), however the placement of students with SEN in different classes and its possible role in class-level differences have not been studied. Therefore, this research focuses on students who are recognized as receiving intensified or special support according to the Finnish learning and schooling support system and on their class placement. The placement is explored by the size of a class, the presence and proportion of other students with SEN and by comparing the placement in special versus regular classes. The main aim in this study could be simplified into the following question: does it matter, what class the student is assigned to?

The aim of the introduction in any scholarly endeavor is to define the topic and describe the context in which the research has been conducted. The idea of this overview is not to go through all the research findings in detail. To some extent, they are described in the original studies. Moreover, not all the readings can be reported here. However, this research has been built on a large and varied body of research and on my own experience in the field of educational research. In this overview, the aim is to set the stage for this study: to provide a context and rationale. Additionally, the aim is to define theoretical and conceptual underpinnings and understanding for why the research questions are posed. Furthermore, the purpose is to reflect critically on the choices I have made during this research process, discuss the main findings, and their possible implications, to define the core concepts as well as to provide future directions. The aim in this introductory part of the study is also to consider the context within both the conceptual and the methodological issues involved in this line of inquiry in general. To some extent, the following chapters are independent of each other and readers can choose the readings based on their own interests and needs.

2 Premises of the study

> "Dear me, half the science of teaching is knowing how much children do for one another, and when to mix them" (Alcott, L. M., 1871, in Little Men).

The main context of this dissertation is the Finnish basic education system and the multi-tiered support model within it. This model, referred to as Learning and Schooling Support, is examined mainly at the class level, by asking questions like where students with SEN are assigned to and how the class placement may affect cognitive and non-cognitive outcomes, both at student and class level. First, I will briefly introduce the multi-tiered support model according to its main features. Then, to set the stage for the current study, I will define the concepts of class size, assigning students to classes and class composition. I will elaborate on how class is defined and contextualized in this study and what is known based on the previous research, mainly from the international perspective. Finally, I will describe the placement of students with SEN in Finland and lastly, discuss student performance in the light of this study.

2.1 Students with SEN in multi-tiered support model

The main purpose of special education is to provide instruction specifically tailored to meet the individual needs of the students who otherwise would not reach the learning goals. To simplify, general education is oriented to the whole age group and special education is directed to specific individuals. In the Finnish multi-tiered system of support, support is provided at three levels, general (Tier 1), intensified (Tier 2), and special (Tier 3) (Basic Education Act 628/1998; FNBE, 2016). This special education system is referred as to Learning and schooling support and it has been based on the three tiers since 2011 (Basic Education Act 628/1998 Amendment 642/2010). From a legislative perspective, all comprehensive school students are educated in the same education system (Basic Education Act 628/1998; FNBE, 2016). From an international perspective, this is in line with the United Nations Convention on the Rights of Persons with Disabilities (2006).

The starting point in the Finnish system is a student's right to learning and schooling support based on individual needs. The main idea of the support provision is preventative, and the purpose is to identify any difficulties early on and provide additional help whenever the student needs it, whether or not a disability has been diagnosed—thus, the system is based on a student's educational needs, not on a medical diagnosis. The overarching idea is to bring

the support services to the student rather than bring the student to the support services (Jahnukainen & Itkonen, 2016). Furthermore, support at all tier levels should be provided immediately when the need arises and there is no need to wait for a specific diagnosis. Consequently, the Finnish support model describes only the delivery of the services, not the actual prevalence of disabilities (Jahnukainen & Itkonen, 2016).

Forms of support include remedial teaching, part-time special needs education, interpretation and assistance services and special aids (FNBE, 2016). Part-time special education is an essential part of the Finnish support system; students at every tier level are entitled to it, without any administrative decisions or diagnoses (Graham & Jahnukainen, 2011). At the Tier 2 level, it is the main form of support (FNBE, 2016; OSF, 2019). Generally, the support methods and tools are almost the same at all tier levels; however, the intensity of the support provided increases from one level to the next (Figure 1) (FNBE, 2016; Thuneberg et al., 2013). Even When planning the support to be provided to the student, it must be taken into account that the need for support may vary from temporary to continuous or from minor to stronger, and that the student may need one or several forms of support (FNBE, 2016).

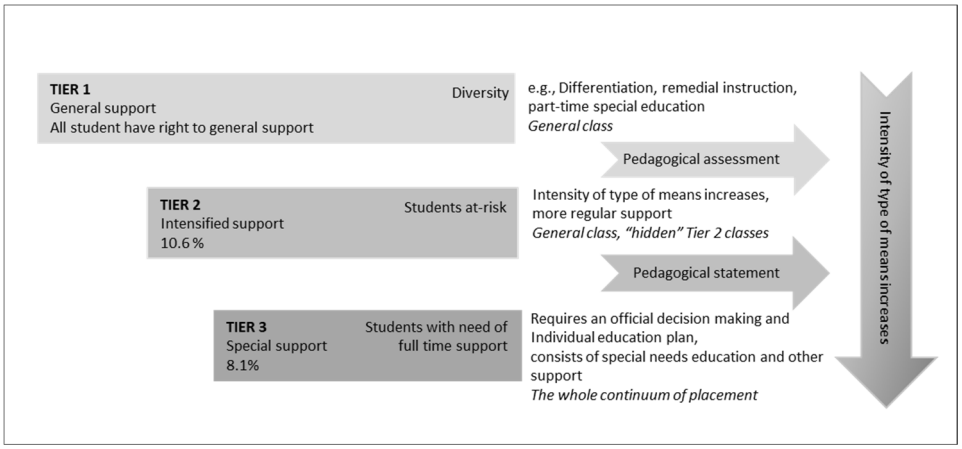

Figure 1 The provision of support in the Finnish multi-tiered support model (FNBE, 2016; OSF, 2019)

Tier 1 general support is applied to all students; it is the first response to a student's need for support and it emphasizes prevention of further difficulties (FNBE, 2016). No specific evaluations or decisions are required. This also means that the practices of delivering Tier 1 support vary notably across the schools. Furthermore, as no official documents are required, there are no official statistics on Tier 1 support. However, an approximation can be given: around 16% of all comprehensive school students received part-time special education at Tier 1 level (Lintuvuori, 2019).

Tier 2 intensified support is for students for whom the primary tier is not sufficient or who are at higher risk. It is implemented when a student needs a longer period of support, multiple support methods simultaneously or more intense support. During Tier 2 support, all forms of support can be used, excluding special needs education provided only at Tier 3 level (FNBE, 2016). The support forms are recorded in a learning plan that must been done according to pedagogical assessment. A total of 10.6% of comprehensive school students received Tier 2 support in 2018 (OSF, 2019).

Tier 3 special support is reserved for students who otherwise cannot adequately achieve their growth, development, or learning goals (FNBE, 2016). This tier concerns 8.1% of all comprehensive school students (OSF, 2019). An official decision concerning Tier 3 support is made by the education provider based on a pedagogical statement, and an individual education plan is drawn up for the student. Special support consists of special needs education and other support needed by the student; in other words, the whole continuum of special education services should be available for the student. Full-time special needs education given by special education teachers can be provided only for Tier 3 students. In addition, the content and the scope of the curriculum can be modified only for Tier 3 students. That is, Tier 3 students study school subjects either according to the general education curriculum (55.1%) or by an individualized syllabus in one or more subjects (45.9%), depending on the severity and nature of the disability (OSF, 2019). The objectives and content of the individualized curriculum are derived from the general curriculum in a way that meets each student's own achievement level, for example, by applying content and teaching materials from the lower grades (FNBE, 2016).

The Finnish support system has many distinctive features and thus, it is not easy to compare it with other educational systems around the world. However, it shares some similarities with other multi-tiered systems of support, such as the Response-to-Intervention (RTI) in the US (Jahnukainen & Itkonen, 2016). Yet, the Finnish system is mainly a framework for structuring and systematizing support, whereas the US RTI is primarily intended for diagnosing and preventing learning disabilities (Björn, Aro, Koponen, Fuchs, & Fuchs, 2015; Jahnukainen & Itkonen, 2016; Sundqvist, Björk-Åman, & Ström, 2019). While the American RTI focuses on measuring the student's responsiveness of the taken actions, the emphasis of the Finnish model is on assessing which means of support are needed by the student.

Multi-tiered systems in general refer to systematic ways to organize support for those who need more intensive instruction, with the aim of identifying difficulties early on (Kauffman, Nelson et al., 2017). Multi-tiered support systems also serve as frameworks for making important educational decisions; in other words, they offer a basis for decisions about students' need for more intensive

instruction or behavioral support and intervention, as well as for evaluating the effectiveness of provided support. This also applies to the Finnish system.

One challenge when describing students with SEN in the Finnish system, at least from the research perspective, is the lack of information on the grounds for receiving Tier 2 and Tier 3 support. The grounds for a decision on special support have not been compiled statistically since 2011 (Lintuvuori, 2019). In a way, this is in line with the international literature, as Kauffman and his colleagues (2017) argue in the *Handbook of Special Education* that the purpose of multi-tiered frameworks is not to classify students, but to make informed decisions regarding the interventions that should be delivered and the resources needed to provide them. Classification as having special educational needs requires extensive mediation between its many consequences both positive—provision of rights and additional resources, and negative—stigmatization and labeling (Richardson & Powell, 2011). Moreover, from a scientific viewpoint, SEN status is admittedly somewhat of an arbitrary label. For example, disability is assumed to be a measurable difference from normal or typical in an individual's ability to accomplish particular tasks (Kauffman, Nelson et al., 2017). The extent that the measured difference deviates from normal is arbitrary. In Finland, students with SEN can be identified by the pedagogical documents drawn up for them according to the National Core Curriculum (FNBE, 2016). Due to the lack of nationwide exact criteria for the support received, determination of the students at different levels of tiers differs between regions and municipalities (Lintuvuori, 2019). Therefore, students in this study have been identified according to the provision of Tier 2 or Tier 3 support as reported by the schools, not based on the research team's independent assessments.

Even in the international research literature, there is no exact, consistent and generally accepted definition for students with SEN. Usually, it covers those for whom a special learning need has been formally identified because they are "failing in school for a wide variety of reasons that are known to be likely to impede a child's optimal progress" (OECD, 2007, p. 18). Furthermore, students with SEN are often those for whom additional resources, in terms of personnel, material or financial, have been provided to support their education (e.g., OECD, 2013). Thus, when considering students with SEN, we are to expect a heterogeneous population. In this study, students who receive Tier 2 or Tier 3 support are referred to as students with SEN or Tier 2 and 3 students and the terms are used interchangeably to designate students who have officially identified needs for their learning and schooling. The terms have been chosen in order to be in line with the journal publishing the original studies. Even though the Finnish definition for students with SEN do not include the diagnoses, and not all Tier 2 or Tier 3 students have actual disabilities, the concept is in line with the International Standard Classification of Education (ISCED) definition of a student with SEN, which states that additional support is provided for individuals who

require it, for a wide range of reasons (UNESCO, 2012). Tier 1 students have been included in the group of students without SEN in the present study.

Disproportionality in the student body is characteristic of students with SEN. It refers to the difference between a given group's proportion of the student population and that group's proportion of students identified for special education without reference to category (Kauffman, Nelson et al., 2017). Disproportionality may involve any identifiable group and may be characterized by over-representation or under-representation. Students with SEN in Finland share similar demographic characteristics to those reported in other studies, such as a higher proportion of boys, lower socioeconomic status, and on average, lower academic achievement (e.g., Hibel, Farkas, & Morgan, 2010; OECD; 2016) Yet, it is important to acknowledge that students with SEN are not automatically low-performing students and low-performing students are not automatically students with SEN (Leino et al., 2018; Smith & Douglas, 2014). Besides, distinguishing between different types of SEN, students with SEN display widely heterogeneous performance profiles. There can be high-achieving individuals among students with SEN, even on average they may perform below average (Lintuvuori, Hienonen, & Hautamäki, 2019).

2.2 Student class assignment in schools

Schooling involves a division of large and diverse student populations, along with other resources, to create arrangements that enable feasible instruction so that teachers can provide students, hopefully gathered in manageable numbers, with appropriate instruction (Dreeben & Barr, 1988). That is, the main mode of learning in schools involves groups of students of the same age interacting with a teacher leading the activity in a limited physical space, directed toward learning a particular topic. To put it more succinctly, students are placed in classes (Ehrenberg, Brewer, Gamoran, & Willms, 2001).

To oversimplify learning in school, teachers teach a certain number of students at a time in a certain physical space, and students learn, some quicker and more easily, some at a slower pace and with more of a struggle. Moreover, there is an underlying assumption of a direct model, in which teaching affects students' achievements and learning in a causal way (Blatchford, Kutnick, Baines, & Galton, 2003). However, teachers do not meet the students out of the context, and the number of students in the classroom as well as many class compositional features can be seen as contextual factors influencing classroom life. They play a part in affecting the behavior of both the students and the teacher. Consequently, both parties will necessarily need to adapt to the classroom context (Blatchford, Baines, Kutnick, & Martin, 2001; Blatchford, Basset, & Brown, 2011).

Schools are large, complex, social organizations that comprise nested layers: student populations are divided by grade, by class, and possibly by temporary

instructional group like small groups of students who need additional help (Harker & Tymms, 2004). Thus, schools operate like many other complex organizations: they divide students into sub-groups that are more homogeneous than the student body as a whole (Gamoran et al., 1995, also Postel, 1937). The aim of the student allocation to classes is to make teachers' work more manageable (Dreeben & Barr, 1988); thus, assigning students to classrooms is not a random process. Furthermore, different classroom compositions and environments might constrain or enhance different organizational and management processes (Wilkinson, Parr, Fung, Hattie, & Townsend, 2002). In general, school effectiveness depends in large part on the ability of the school to respond to individual student needs, by assigning students to sub-groups that can enhance their learning (Rea, McLaughlin, & Walter-Thomas, 2002).

All students face a school-based decision of placement—a placement that can vary in terms of the number of other students in class and in terms of the composition of the class. That is, students in the same classroom differ in prior knowledge, readiness to learn, motivation, socioeconomic background, gender and in their need for support learning and schooling. Therefore, the class composition consists of the background of students, the average performance level of the class, overall learning motivation, classroom climate etc. As prior studies have shown, the performance differences in Finland lie between classes rather than between schools, it is justified to assume, that class compositional effects could partly explain these between-class differences. As students are not randomly assigned to classes, the achievement differences are found to be related to differences in the composition of the body of students—in classes at schools (Harker & Tymms, 2004). To simplify this, students with a similar initial level, but who are placed in different classes, can be predicted to have different achievements depending on the average achievement level of their classmates, which in turn could lead to the conclusion that class placement matters.

2.2.1 Class size

> "If children are not learning as much as they are expected to, class size is a convenient scapegoat" (Insel & Lindgren, 1978).

A defining characteristic of any class is the number of students for which a teacher holds responsibility for instructing. The number students in a class can vary, and increasingly, so can the number of teaching staff. In short, class size refers to the actual number of students taught by a teacher at a particular time, and the number of students in a class has the potential to affect how much is learned in a number of different ways. The general assumption often is that, other things being equal, smaller classes will enable teachers to provide a better quality of instruction for students. In line with this, research on class size has tended to assume an underlying direct and causal model—the focus has been on the effect of the

number of students in class on student performance and development. Yet, schools are complex places, impacted by many factors (Harker & Tymms, 2004). This poses challenges to class size research. The aim here is not to go through all the large body of class size literature, but rather to focus to draw some conclusions on the point of view of students who, more than others, need support for their everyday learning. Before that, the term class size is defined and discussed in the context of Finnish schools.

Class size is often used as an indicator of education in international comparisons. The challenge is that it can be defined in several ways. One option is to use student-teacher ratios. It is defined as the number of students in the school divided by the number of full-time teachers for an entire school (e.g., OECD, 2019a). This ratio often includes principals and special education teachers, and it is argued to have very little to do with actual class size (Akerhielm, 1995). The other option is to calculate class size by dividing the number of students enrolled by the number of classes (OECD, 2019a). These two indicators measure very different characteristics of the educational system. While the student-teacher ratio provides information on the level of teaching resources available, class size measures the average number of students that are grouped together in classrooms (OECD, 2019a). Konstantopuolos and Traynor (2014) further argue that in many studies, class size is not measured accurately because data about the actual class size in each classroom are not available (instead, it is calculated as administrative reported enrolment divided by number of group teachers). This has led to the point that class size is represented by average class size. However, the average number of students in class does not necessarily explain about the actual class size. In terms of student learning, what matters is the number of students who are present in the same physical space interacting among themselves and with their teacher (Ehrenberg et al., 2001). Still, in many class-size studies, the correlation between the actual class size and the calculated one has been quite high (Hoxby, 2000). In the present study, the measure for class size was drawn from the student lists received from the education providers; hence, class size was defined as the number of students assigned to a certain class, like 7B or 9D. Naturally, it must be acknowledged that class size is not a fixed measure. The number of students present in the class at any time may be slightly different from the number in the class register; thus, it can vary to some extent during a school day and a school year.

Class sizes at the primary and the lower secondary education level are slightly different constructs. Class size is slightly more straightforward in primary education, as it can be defined simply as the number of students assigned to a classroom teacher (e.g., Hoxby, 2000; Molnar et al., 1999). Furthermore, primary school students spend much of their school time in a single classroom with a regular group of peers and a single teacher who instructs them in several subjects. In the subject teacher-based system of lower secondary education, class size is

more difficult to identify, since class size tends to vary by subject matter. Naturally, students receive instruction in classrooms, but not in the same class all day, and they are exposed to several teachers and changing groups of classmates. However, to some extent, classes form meaningful and stable groups because most students stay together as a group when different subjects are taught by different teachers.

What kind of numbers do we mean when we talk about class sizes? Most of the research is based on classes in the Western literature, i.e., European and North-American, where the average class sizes vary from 20 to 30 students, whereas in many non-Western countries the typical class size well exceeds this, and typical class sizes are more like 30–40 (Hattie, 2005, OEDC, 2014; OECD, 2019a). In Finland, class size is not part of the annual, national statistical compilation. However, since 2008, the information on class sizes has been collected as part of the national collection of teacher data (Opettajatiedonkeruu) carried by Statistics Finland every two to three years. The average class size in primary education has varied from 20.7 students in 2008 to 20.1 students in 2016, and in lower secondary education between 17.3 and 15.9, respectively (Karjalainen & Lamberg, 2017). The class size tends to increase during primary education from 18.8 students in the first grade to 22.5 student in the sixth grade. The class size in the lower secondary grades is smaller than in primary education and the lower secondary class size is quite stable from seventh to ninth grade. The national averages naturally mask all the variation in class sizes and for example, the numbers do not report the size of classes in which students with SEN are placed.

Particular interest in Finland has been paid to the ratio of classes exceeding 30 students, and the trend has been decreasing. The proportion of these classes in 2008 was 2.4% of all classes, and in 2016, it was only 1.0% (Karjalainen & Lamberg, 2017). From an international perspective, Finnish class sizes are slightly lower in comparison with other OECD countries. In primary education and lower secondary education, the average in Finland is 20 students[1] (OECD, 2019a). The averages across OECD countries are 21 and 23, respectively. In respect to the student-teacher ratio, the picture is slightly different: the student-teacher ratio in primary education in Finland is 13, and in lower secondary education 9, whereas the OECD averages are 15 and 13, respectively (OECD, 2019a).

Above all, class size is a highly political topic in Finland and globally, not least due to the fact that it is often directly related to the current economic situation. Policymakers, different stakeholders, teachers, and parents are naturally interested in identifying learning environments that can increase academic development and that can draw out students' full potential. Furthermore, it is clear that policy-level decisions affect educational practices directly through municipal and school

[1] Class size as calculated by the OECD is different from that calculated by Statistics Finland in the national collection of teacher data. The two ratios are not comparable.

financing, for example class sizes. In general, class size reduction can be an appealing school intervention because it is considered to be easy to implement (e.g., Ehrenberg et al., 2001; Hattie, 2005; Hoxby, 2000). Thus, there is constant pressure to reduce class sizes or at least prevent them from increasing. However, even though class size reduction is seen as an easy school reform to implement, it is never done without a cost. As was said, class size is always bound to economics (Hanushek, 2003). Hoxby (2000) translates it into an education production function—the assumption that there should be a relationship between the input (reduced class size) and output (increased development of performance). This aligns with the idea that increasing education funds will automatically mean better results and superior educational outcomes. Reducing class size leads to an increase in the number of groups and thus, in the number of teachers needed. Teacher salaries comprise 80% of the expenditure in education (Juva, 2008). In fact, class size reduction is seen as the most expensive policy reform (Hanushek, 1999; Hattie, 2005; Schanzenbach, 2010; Yeah, 2009).

The Finnish government has provided considerable amounts of extra funding for class size reduction. The funding was first released in 2009, and it continued until 2015. At first, the funding was not "aimed at special education" (Ministry of Education and Culture, 2011). Later, the emphasis was on mainstream classes with students with SEN (Ministry of Education and Culture, 2012). In recent years, the funding for class size reduction has been embedded in the discretionary subsidies for promoting equity in pre-primary and basic education (e.g., Ministry of Education and Culture, 2017).

The basic education legislation does not stipulate the class size for regular classes in Finland. Every now and then it is suggested and even demanded that the limits for the number of students in class be stipulated. However, the education provider's right to make the student allocation decisions has been preferred (Lahtinen & Lankinen, 2015, p. 224). The Basic Education Act (628/1998) only states that "the teaching groups shall be formed so that the instruction can achieve the objectives set in the curriculum". In other words, schools allocate students to classes of different sizes in the way they find best. Within-school variation in class size is seldom random, as it seems that even when there is the possibility to organize classes of equal and moderate sizes, schools allocate students to classes of varying sizes (Kupiainen & Hienonen, 2016, p. 91; also Schanzenbach, 2010).

2.2.2 The why's of class size

Class size is one of the longer standing and debated question in educational research (Blatchford & Russell, 2019). One of the main reasons for the countless class size studies is their ambiguous and contradictory findings that, in turn, have challenged researchers around the world to attempt to solve the eternal question.

Given the vast body of research literature, the aim in this chapter is to focus on class size from the viewpoint of students with SEN.

If one result is to be drawn from the vast body of class size literature, it may well be that class size and student-teacher ratio matter more to some groups of students than to others. There is compelling evidence that lower-performing and disadvantaged students could benefit from smaller classes more than others (e.g., Blatchford, Basset et al., 2011; Blatchford, Goldstein, Martin, & Browne, 2002; Blatchford & Mortimore, 1994; Finn & Achilles, 1990; Finn & Achilles, 1999; Hargreaves, Galton, & Bell, 1998; Molnar et al., 1999; Schanzenbach, 2010). However, contradictory findings have also been found, as in some studies the effects of class size reduction have been more pronounced in classes of higher-ability students (Hoxby, 2000; Konstantopoulos & Traynor, 2014; Rice, 1999).

Generally, both common sense and pure logic suggest that with more students in the class there will be more potential for distraction, and less time for student-teacher interaction and individual instruction (e.g., Blatchford, Edmonds, & Martin, 2003). Conversely, in small classes teachers have more opportunities to engage students and keep them on task; more time for individual, one-to-one instruction; and greater knowledge of their students, and students have better knowledge of their classmates. Furthermore, students are more likely to interact with their teachers, there are fewer discipline problems, and there is more time for helping students to acquire common content or skills (e.g., Blatchford, Bassett, & Brown, 2005; Blatchford, Basset, Goldstein, & Martin, 2003; Blatchford, Basset et al., 2011; Finn & Achilles, 1990; Glass & Smith, 1980; Molnar et al., 1999). In sum, in small classes students have more opportunities for individual attention, whereas; in large classes children are more likely to be one of the crowd. However, there are also indications that teachers use the students as the audience in smaller classes (Hargreaves et al., 1998), and that students in small and large classes spend the same amount of time on and off tasks (Blatchford, Bassett et al., 2005).

One of the main challenges in studies investigating class size is that the allocation of students to different classes is not a random process (e.g., Harker & Tymms, 2004; Hoxby, 2000; Konstantopoulos & Traynor, 2014; Kupiainen & Hienonen, 2016; Paufler & Amrein-Beardsley, 2013). For example, in Finland, students with SEN are disproportionately found overrepresented in smaller classes and, at times, higher performing students in larger proportions in selective classes[2] (Kupiainen & Hienonen, 2016; Kupiainen, 2019). The same phenomenon is recognized elsewhere as well (e.g., Akerhielm, 1995; Pedder, 2006). Consequently, any positive effect of a small class on student performance may be concealed (Dobbelsteen, Levin, & Oosterbeek, 2002). To cut a long story short, the student composition of the class should always be studied alongside the

[2] The term selective class (painotetun opetuksen luokka) refers to a class with a special emphasis (e.g. music, science). The student admission is based on application and selection via aptitude tests in the emphasized subject area (Kosunen, 2016).

number of students in class. However, before moving on to discuss the compositional effect, I will shortly address the complexities of class size effects.

Despite the logical reasoning, there is a lack of systematic evidence of class size reduction benefits. Drawing on the well-known, cited and also criticized analysis of meta-analyses of John Hattie (2005), he summarizes the why's of class size research in one question: Why hasn't reducing class size led to major improvements in student learning? One explanation can be that many factors other than the class size influence more what and how much students learn (Ehrenberg et al., 2001; Hattie, 2009). Another explanation of neutral effects can be the teacher effect. It may be that the effects of smaller class sizes depend greatly on teachers altering the way that they teach (Ehrenberg et al., 2001). Findings from several studies have indicated that teachers tend to believe that class size has a major effect on what they do, and on the effectiveness of what they do (Hargreaves et al., 1998; Pedder, 2006). There are claims, that teachers do not vary their teaching according to the number of students in class (e.g., Betts & Shkolnik, 1999; Hattie, 2005; Slavin, 1989). However, even if a teacher does not teach differently in a smaller class, a teacher can devote more attention to each student during every teaching activity that has an individual element (Hoxby, 2000). Furthermore, naturally, the effect of class size cannot be accounted for entirely by the effect it has on teaching practices (Bourke, 1986). It can also be the case that there may be enhanced opportunities for learning in smaller classes when teachers act in certain ways, but students may not always have developed the skills or attitudes to take advantage of these opportunities (Pedder, 2006).

To date, the size of a class has been studied mainly in terms of regular classes. Nevertheless, smaller class sizes for special education classes and individualized instruction have been identified as an important factor for meeting the needs of students with special needs (Zarghami & Schnellert, 2004). Yet, to date, few studies have addressed this question. Indeed, it is clear that many of the benefits smaller classes are thought to have could be especially beneficial for students with SEN—namely, more individual attention from their teachers, fewer discipline problems, greater flexibility in teaching strategies, more feedback on students' work and greater teacher knowledge (Blatchford, Basset et al., 2011; Bourke, 1986; Finn & Achilles, 1990; Molnar et al., 1999). In short, more individualized interaction between teachers and students affects the students' learning and motivation (Blatchford, Bassett et al., 2005; Blatchford, Moriarty, Edmonds, & Martin, 2002; Harfitt & Tsui, 2015). Furthermore, there is evidence that low-performing students benefit more from individualized instruction than high-performing students (van Hek, Kraaykamp, & Pelzer, 2017).

2.2.3 Class composition

> "When I ask teachers if they would choose between a class size of 15 when I choose the students, or a reduction of five from their current class and they choose the students, they nearly always prefer the latter" (Hattie, 2005, p. 416).

One of the main problems in class size research has been that most research has treated the classroom as a black box with the assumption and expectation that any effects of class size on student performance would be automatic and direct. It is clear that the number of students in the class necessarily affects what a teacher can do (Blatchford, Baines et al., 2001), yet the picture is more complex. As the quote at the beginning of the chapter points out, the number of students in class is only one factor influencing what happens in the classrom.

Much of class composition studies deal with tracking and ability grouping. Finnish basic education does not operate officially with tracking and the focus of this study is on the processes of assigning students, which may produce a kind of informal tracking. Tracking and ability grouping in general are intended to create homogeneous learning groups to adapt the instruction to the needs of the specific group of students by dividing students purposely for instruction according to their assumed capacities for learning. (e.g., Hanushek & Wößmann, 2006; OECD, 2012; Slavin, 1990). This practice is in line with the thought that teaching a homogenous group of students could be more efficient—it would allow teachers to tailor their instructional approaches and to find the most appropriate level and pace of instruction (Belfi, Goos, De Fraine, & Van Damme, 2012; Gamoran et al., 1995). For higher performing students, this is done to maintain the interest when the goals for learning are high enough, whereas for lower performing students, this is done to encourage them to try their best, without fear of failure and comparison to higher achieving students. Thus, tracking has been justified by a better promotion of all students according to their achievement potential and by providing the best possibilities for their development (Slavin, 1990). Yet, this kind of grouping is also criticized. Although it can be an attempt to provide appropriate instruction for different groups of students, in practice low-performing students are placed in the lower tracks and they may end up receiving inferior instruction compared to their higher-track peers (Gamoran, et al., 1995; Hattie, 2002). There are indications that belonging to a high ability class positively influences students' academic achievement, whereas the opposite is true for belonging to a class with a large proportion of lower-performing students (Peetsma, van der Veen, Koopman, & van Schooten, 2006; Van de gaer, Pustjens, Van Damme, & De Munter, 2006).

Even though Finnish students are allocated to classes in a more equal manner, without explicit ability-grouping, it does not mean that the placement decisions are done randomly. Students are allocated to classes mainly based on principals'

decisions on establishing and composing classes. Decisions on the best allocation of school resources include decisions about assigning both teachers and students to classrooms. Typically, such decisions involve determining the optimal number and composition of students in a classroom in order to maximize student learning. There has been little research on what affects student allocation in schools. A North American study on purposeful and random assignment of students into classrooms indicated that prior academic achievement, special education needs, giftedness, and gender heavily influence the placement decisions, in addition to behavioral issues and needs, and prior grades (Paufler & Amrein-Beardsley, 2013). At the lower secondary education level, in a Finnish study with a national sample and principal survey indicated that when making placement decisions, prior knowledge on students (if available), peer relations, support needs, and subject-matter choices such as language are considered. Furthermore, aims to create balanced classrooms were reported in terms of gender division and prior performance (Kupiainen, 2019, p. 94). In addition, at least in the larger cities, there are selective classes with a special emphasis and they cover both academic (e.g., science and mathematics) and non-academic (e.g. music and sports) subjects (Koivuhovi, Vainikainen, Kalalahti, & Niemivirta, 2017; Kosunen, 2016).

The aim of assigning students can be to create balanced, heterogeneous classrooms, with an effort to create classrooms in which the composition of the class is representative of the school. The other option is to aim at creating more homogenous classrooms with an intention to reduce the heterogeneity of instructional groups. These two different practices can result in different outcomes. In general, students in classes that are heterogeneous, in terms of the ability levels of the students, may learn more, or less, than students enrolled in classes in which students are homogeneous in terms of their ability levels. (Ehrenberg et al., 2001). There are indications that learning in homogenous classes has certain advantages (Hoxby & Weingarth, 2005). There is evidence that higher-performing students often benefit more from learning in homogeneous classes. Evidence also suggests that learning in heterogeneous classes has more advantages for students with low or medium ability (Kuzmina & Ivanova, 2017). However, other studies have not confirmed the positive effects of homogenous classes (Slavin, 1990) or any effect at all (Hanushek et al., 2002).

In many school effects studies, achievement differences are found to be related to differences in the composition of the student body (Harker & Tymms, 2004; Reynolds et al., 2014). This is known as the *compositional* effect. A compositional effect—also referred to as a contextual effect—in a statistical sense can be described as "an effect of a school, class, or other group level aggregate of an individual level variable over and above the effect of the same individual level variable on a certain outcome variable" (Harker & Tymms, 2004, Van de gaer et al., 2006; Televantou et al., 2015). The compositional effect can be understood with the following example. It might be expected that a student will make more

progress if the average achievement level of the class is higher, and, conversely, less progress if the average achievement level of the class is lower (Blatchford, Goldstein et al., 2002). Alternatively, according to Gamoran and his colleagues (1995); students perform better in schools primarily composed of high ability students than in schools primarily composed of low ability students, after controlling for the students' own abilities.

Following Harker and Tymms (2004), the term compositional effect has been used in this study to describe the statistical estimate of the additional effect obtained by the aggregated variable at the class level, over-and-above the variable's effect at the student level. It means that a class-level aggregate of a student-level variable is hypothesized to make an independent contribution to the explanation of outcome variance. In other words, does the classroom composition affect the achievement of an individual student (Zimmer & Toma, 2000)? The methodological features are represented more thoroughly in Chapter 3.5.1. Next, more theoretical aspects of the class composition effect are described.

2.2.4 What shapes learning in classrooms?

Students' learning and performance is influenced by their personal characteristics. In addition, students' learning is strongly influenced by the educational context in which it occurs (i.e. schools and classrooms). The classroom context is defined by students' classmates with whom they experience teaching and learning, the peers with whom they choose to interact, and the teachers who instruct them. Students take their cues for expectations for appropriate behavior from the individuals with whom they interact in schools, which means not only fellow students but teachers as well (Dreeben & Barr, 1988). Thus, from the viewpoint of an individual student, there are two main ways the classroom processes affect their learning—what the teacher does in the class and what other students do in the class. In their eminent work of school compositional effects on academic performance, Harker and Tymms (2004) have grouped the effects into four main categories: peer effects, teaching effects, facilities effects, and phantom effects. The first two can be applied in class composition effects as well as the last one. Though both the peer and teacher effect as such are out of the scope of the present study, they are briefly described here as they can be seen as a relevant part of extensive explanations for the class composition effect. The phantom effect is discussed in Chapter 4.3.

As has become evident, classes in schools represent different compositions of students which affects students' learning. However, learning is also heavily influenced by what and how students are taught. Even though teachers are not presented as a variable in this study, it must be noted that teachers bear the primary responsibility for shaping students' learning experiences and may have more impact on student achievement than any other school-based factor (Rivkin, Hanushek, & Kain, 2005; Wilkinson et al., 2002). The classroom composition

affects the teaching, namely different teaching techniques, disciplinary procedures, teacher commitment, and classroom climate. Thus, different instructional activities and materials in different classes are expected to be found (e.g., Gamoran et al., 1995; Harker & Tymms, 2004; Thrupp, 1995). The teacher influence is not on the same for all students. It has been found that qualified teachers were especially beneficial for lower-performing students (Nye, Konstantopoulos, & Hedges, 2004). In addition, there are claims that teachers do not always change their instructional methods according to classroom composition, but they may change the pace and the materials provided to students (Wilkinson et al., 2002). For example, in classes comprising higher performing students, teachers tend to use more complex tasks and autonomy-supported learning whereas students in homogenously grouped low-achieving classes often have lower expectations of teachers (Kuzmina & Ivanova, 2017, see also Snow, 1989).

Students are affected by teachers, and teachers are affected by students in the class. In the largest international survey of teachers, the Teaching and Learning International Survey (TALIS), teachers' self-efficacy and job satisfaction is investigated in relation to classroom environment. Interestingly, class size had a minimal effect on either teaching efficacy or job satisfaction in a few countries (OECD, 2013, p. 190). Moreover, TALIS data indicated that it is not the number of students but the type of students which has the largest association with teachers' self-efficacy and job satisfaction. Certain classroom characteristics can make a teacher's work more challenging. In TALIS, student composition in the classroom is characterized by low academic achievers, students with behavioral problems and academically gifted students (OECD, 2013, p. 193, 198). According to the survey, classrooms were considered to be challenging if more than 10% of students in the classroom were low academic achievers or more than 10% of students had behavioral problems. However, it was not the percentage of these students as such that influenced directly on a teacher's self-efficacy or job satisfaction. Rather, it was the time the teacher spends dealing with the classroom-management issues that these students—or other students in these classes—may cause.

Besides teachers, students' learning experiences depend greatly on their fellow students in the classroom. In class, each student is surrounded by classmates who represent a certain range of academic competencies, history and different backgrounds. Composition of a classroom—that is, the characteristics of the students in the class—affect the educational achievement of an individual student. This influence of the students in a classroom is often referred to as a peer effect (Zimmer & Toma, 2000). According to Gamoran and his colleagues (1995), the intellectual capacities of classmates constitute an important classroom resource, and according to Dreeben and Barr (1988), class composition brings about the normative influence on how to behave. It is clear that students use their classmates

as a normative reference group (Wilkinson et al., 2000). From the sociocultural perspective, social interaction with more skilled peers can facilitate a child's cognitive and academic development (Vygotsky, 1978). In processes of social comparison and socialization, students can internalize the values and norms of their classmates (Van de gaer et al., 2006). This means that beliefs about the self in class are formed in comparison to others. That is, students tend to compare themselves with their classmates who have slightly better achievements than their own. This perception can have a negative effect on the students' self-concept, educational expectation or achievement, which is discussed as the Big-Fish-Little-Pond-Effect (BFLPE) (Marsh, 1987; see also Huguet et al., 2009). In short, the BFLPE refers to a model in which class-level average ability is negatively associated with students' academic self-concepts (Marsh & Parker, 1984). Equally performing students have a lower academic self-concept in high-ability classes than in low-ability classes because in higher-performing classes they compare themselves with other high-achieving students and thus have a lower self-perception of their own abilities. However, when a student is placed in a lower-performing class, self-concept is higher because there is no detrimental comparison with high achievers (Marsh, 1987; Marsh & Parker, 1984; see also Dreeben & Barr, 1988). There is also evidence on reciprocal effects between academic self-concept, motivation and student performance (Gorges, Neumann, Wild, Stranghöner, & Lütje-Klose, 2018).

Furthermore, the social context of the class provides opportunities for individuals to imitate each other and to learn from more abled peers. There are indications that the peer effect is different on differently performing students. In a study by Zimmer and Toma (2000), the effects of peers appeared to be greater on low-performing students than on high-performing students. In other words, lower-performing students benefited more from their higher ability of the peers. However, it cannot be assumed that what is learned from others is always positive (Kauffman & Pullen, 1996). Peers can offer both desirable and undesirable models for learning and behavior in every class. In addition, it must be kept in mind that the compositional and peer effects are not the same. Peer effects probably occur within small clusters of students (Wilkinson et al., 2002), and as said, are not the focus of this study.

In a nutshell, classroom composition effects develop from a bidirectional process: students react to classroom structure, climate and to their peers, and classes in turn react to the composition of the student body (Harker & Tymms 2004). A class is a sum of the students and the teacher in it.

2.3 The placement of students with SEN

One crucial aspect of student allocation is the placement of students with SEN. Class placement in this study refers to the type of the classroom where students

with SEN receive instruction. While students without any specific support needs are placed in classes of slightly different sizes and peer composition, students with SEN face a placement decision of a wider range. The placement can be anything from a full-time placement in a regular class to a full-time placement in a special class in a special school. One of the dividers is whether a student is mainly taught in a regular class by a classroom or a subject teacher, or in a special class by a special education teacher (Zigmond & Kloo, 2017). Generally, there is disagreement about the desirability of various placement options and there is still a lack of objective scientific investigation on this matter and thus, many of the different placement practices do not rest clearly on the research-based facts (Kauffman, Nelson et al., 2017). Therefore, the placement choices depend on the severity and nature of the need for support, municipal and school level decisions and practices, available resources and ideological aspects.

The placement of students with SEN is a timely topic that evokes strong opinions and heated debates. The question is many-sided and it includes at least the following points: 1) where students with SEN are placed, 2) what are the reasons for the different placement decisions, 3) how does the placement affect the students with SEN, and, 4) how does the placement of students with SEN affect the other students. The first and the second points are described and discussed briefly in the Finnish context in this chapter. An aim of Study II was to respond to the third and fourth points, and Study III to the third.

Students with SEN can be placed in either special or regular classes and the placement can be either full- or part-time. The main question thus is, are students learning equally in both settings and are there differences favoring one setting over the other? The question has not been resolved yet. One reason for this is that these themes have been the subject of little objective, scholarly investigation and thus, many of the different placement practices do not rest clearly on research-based facts (Kauffman, Nelson et al., 2017). The growing tendency is to organize the learning and support in regular classes, yet, there is still a belief that the needs of a certain group of students are better met in special classes (Fuchs, Fuchs, McMaster, & Lemons, 2018; Kauffman Nelson et al., 2017; Zigmond & Kloo, 2017).

The topic has been a subject of heated discussion and polarized views. Advocates of the placement in regular classes perceive it as a matter of a legal justice, and as a human and social value. They also argue that separate classes are stigmatizing and exclusionary (Kauffman, Nelson et al., 2017). Thus, supporters of the regular class placement suggest that students with SEN in regular schools and classes can feel more appreciated as capable learners (Bakker & Bosman, 2003). On the contrary, proponents of special classes argue that these settings can provide a more protected learning environment for students with SEN (Kojac, Kuhl, Jansen, Pant, & Stanat, 2018). It means more adapted instruction, individualized feedback, and a less competitive climate in the classroom

(Peetsma, Vergeer, Roeleveld, & Karsten, 2001). Regular class placement can also be seen as a cost-cutting effort and cost cutting by the education provider. In the worst case scenario, in teachers' views, it can be seen to lead to situations in which "everybody suffers" (Honkasilta et al., 2019, p. 490). On the one hand, there are concerns that without adequate resources, the needs of students with SEN cannot be met in regular classes. On the other hand, if the insufficient resources are directed towards the struggling students, it can lead to neglecting the other end in the class, the more advanced students. However, according to the literature, support provided in regular classes could benefit all students and thus, create positive spillover effects (Hanushek, Kain, & Rivkin, 2002; Keslair, Maurin, & McNally, 2012; Thuneberg et al., 2013). In line with global policies, whatever the arguments for and against the placement of students with SEN are, the placement in regular class is seen as every student's right (UNESCO, 1994). Yet, there are also opposing views, for example among the principals who call for individually defined placement. They have argued that it is also every student's right not to get integrated (Jahnukainen, 2015). Furthermore, the Trade Union of Education in Finland (2019) has recently called for Tier 3 students' right to special class placement to be stipulated by law.

In the era of inclusive education, we still have special classes and special schools. Following this, one question is, what makes the special classes different from the regular class? The general idea is that at least in theory, special education and support can be provided anywhere (Kauffman & Pullen, 1996) and the placement itself do not enhance or impede learning and support. However, Zigmond and Kloo (2017) state that we have come to believe that special education is so not so special that it can be delivered in a regular classroom by a regular teacher for a large group of students. Furthermore, they also state we have forgotten that special education is supposed to be special and that wherever it is delivered, it is supposed to be different. Furthermore, there are also concerns that the capacity of general education to provide adequate support for students with SEN has been overestimated (Fuchs et al., 2018).

One explanation for the specialness of special classes is the teacher. While class and subject teachers are prepared to teach content and curricula to the large groups of students, the special education teachers are specifically prepared to apply pedagogical skills and instruction to teaching individuals or small groups of students with specific learning needs (Zigmond & Kloo, 2017). Even as the placement in regular classes is increasing, the special education teacher and the general education teacher have different job descriptions. The role of the special education teacher is to teach what cannot be learned elsewhere—it is, by a definition, special teaching (Zigmond & Kloo, 2017). This is also supported by the finding that qualified teachers were especially beneficial for students from disadvantaged backgrounds (Nye et al., 2004). The Finnish legislation states that the special needs education given students at Tier 3 level is provided by a special

education teacher (Basic Education Decree 852/1998, 1§). In addition, there are qualification requirements for teaching personnel for part-time special education, for special needs education given in certain subjects and in special classes (Teaching Qualifications Decree 986/1998). When the instruction is given in conjunction with mainstream education, class and subject teachers are entitled to give instruction to Tier 3 students in subjects in which their individual education plans (IEP) do not require special needs education (FNBE, 2019).

The second explanation is the number of students in a special class. By law, the teaching group for Tier 3 students may consist of a maximum of ten Tier 3 students, with some exceptions (Basic Education Decree, 852/1998, 2§). In short, the number of students in a class can affect how much time the teacher is able to focus on individual students and their specific needs rather than on the group as a whole. This has been discussed more in Chapter 2.2.2.

The third argument rests on the class composition. Smaller special education classes may be more homogenous in student body than the larger regular classes, that is, students with SEN in special classes can receive instruction together with classmates with similar difficulties and performance level. This can lead to stronger feelings of relatedness and increase their motivation (Bakker, Denessen, Bosman, Krijger, & Bouts, 2007). In general, research on placement effect on psychosocial outcomes has favored special class placement. Students with SEN in special classes or in special schools are found to have higher self-perception and higher academic self-concepts than students with SEN in regular classes and schools (e.g., Bakker & Bosman, 2003; Bear, Minke, & Manning, 2002; Belfi et al., 2012; Kocaj, Kuhl, Kroth, Pant, & Stanat, 2014; Törmänen & Roebers, 2017). The main explanation is that students in special classes compare themselves to students with a similar performance level, which can lead to a more positive self-perception concerning school tasks. And the other way around, students with SEN in regular classes might become less motivated and self-confident when they compare themselves to their higher-achieving peers (Ruijs & Peetsma, 2009). However, contradictory views exist. Elbaum (2002) found no consistent relationship between the placement and academic self-concept of students with SEN. However, the same study found that motivational outcomes tend to favor special classes.

With regard to cognitive outcomes, however, the previous findings suggest that students with SEN would benefit more from a placement in regular classes when different measures of academic outcomes have been studied (e.g., Dessemontet, Bless, & Morin, 2012; Kojac et al., 2018; Peetsma et al., 2001; Rea et al., 2002). There has also been research that has shown mainly neutral effects of regular class placement (Fore, Hagan-Burke, Burke, Boon, & Smith, 2008; Hanushek et al., 2002; Kalambouka, Farrel, Dyson, & Kaplan, 2007; Ruijs, 2017, Ruijs, van der Veen, & Peetsma, 2010; Scharenberg, Rollet, & Bos, 2019). Three meta-analyses on the topic, Carlberg and Kavale (1980), Wang and Baker (1985), and Oh-Young

and Filler (2015) from different time periods have reached the same conclusion when considering academic outcomes: separate settings are not as beneficial as are the more integrated settings. Most students with SEN in more integrated settings outperform those in less integrated settings on academic outcome measures (Carlberg & Kavale, 1980; Oh-Young & Filler, 2015; Wang & Baker, 1985). Some explanations can be presented. Students with SEN may achieve higher because they can learn from more able students in regular classes (Ruijs & Peetsma, 2009). There are also suggestions that students who observe the performance of slightly higher-achieving students are likely to exhibit an improvement in academic outcomes (Huguet, Dumas, Monteil, & Genestoux, 2001). In addition, it can be argued that when students with SEN are placed in regular classes, they can feel that they belong to a group that is positively valued; and thus, students with SEN in regular classes may have higher achievement motivation through basking in reflected glory of the perceived accomplishments of their peers (Dockx, De Fraine, & Vandecandelaere, 2019; Marsh, Kong, & Hau, 2000).

Finally, the question is, who is placed in a regular and who in a special class. A thorough investigation of this question is beyond the scope of this study, but previous research has suggested that certain groups of students (i.e., high-incidence disabilities) are more likely to be placed in regular classrooms whereas students with significant disabilities are more likely to be placed in separate settings (Morningstar, Kurth, & Johnson, 2017; Oh-Young & Filler, 2015).

In the present study, classes are characterized as regular and special classes even though it is not completely in line with the current ideology to avoid the classification. In the Finnish legislation, the definition "in conjunction with other instruction" refers to the regular classes (Basic education Act 628/1998, 17§), and in the National Core Curriculum the term "a mainstream education class" is mentioned once (FNBE, 2016, p. 72). In the international literature, the terms general class and regular class are also used. To make the text more readable, classes which consist only of Tier 3 students and when the number of students does not exceed the statutory class size maximum of 10 students, are referred to as special classes. All the other classes are called regular classes as this was the most often-used term in the international research cited in this study. These regular classes may consist of only students without defined support needs or students without defined support needs and students at Tier 1, 2 and 3 levels.

As education providers, Finnish municipalities handle the practical teaching arrangements and are responsible for the effectiveness and the quality of education in their districts (FNBE, 2016). There are no regulations governing mainstream education class size and schools determine how to assign students to classrooms (Lahtinen & Lankinen, 2015, p. 224–225). The Basic Education legislation sets some guidelines for the placement of students with SEN, but they leave room for education-provider-based decisions. The prevailing view in Finland and globally

is that all students must have access to and should be served in regular classes and schools, and schools should accommodate students and meet their needs (UNESCO, 1994). In line with this, according to the National Core Curriculum, support at all tier levels should be provided in a student's own teaching group and school by means of various and flexible arrangements, unless the student's best interest requires otherwise (FNBE, 2016). What constitutes *student's own teaching group* is not clear, however, it can be assumed that it is in an age-appropriate class a student would be assigned to without any support needs.

Tier 1 and Tier 2 support should be provided as a part of mainstream education in regular classes. The basic education legislation does not stipulate the placement of Tier 2 students, nor did the previous National Core Curriculum (FNBE, 2004). This has apparently led to various interpretations across the education providers. Consequently, in the recent National Core Curriculum from 2014, it was specified that Tier 2 level support is provided as a part of mainstream education using flexible teaching arrangements (FNBE, 2016). Yet, the mention of the flexible arrangements is still open to the interpretations. Since the normative assumption for Tier 2 students' placement is in regular class, the information on the placement is not covered by the national statistics. Previous research has revealed that some schools manage their student population by grouping Tier 2 students into classrooms containing only Tier 2 students or containing only Tier 2 and Tier 3 students (Hienonen & Lintuvuori, 2019; see also Kupiainen & Hienonen 2016; Lintuvuori, 2015). Some of these first mentioned classes were explained by the practice of flexible basic education (JOPO, joustava perusopetus). These Tier 2 classes are not a common practice, but are an existing one, and this forms a hidden structure within the system. In general, the discrepancy between the legal rulings and actual school practice are likely to partly explain the contradictory research findings and interpretations regarding placement effects for students with SEN (Huber, Rosenfeld, & Fiorello, 2001).

Tier 3 support is provided "in conjunction with other instruction or partly or totally in a special education classroom or some other appropriate facility" (Basic Education Act 628/1998, 17§), and by law, it is always stated in the decision on special support. While the size of regular classes is not regulated by law, the teaching group for Tier 3 students must usually consist of a maximum of ten Tier 3 students. The class size maximum for students within prolonged compulsory schooling is eight students and, for students with profound developmental disabilities it is six students (Basic Education Decree, 852/1998, 2§).

Of Tier 3 students, 21.3% studied in regular classes full-time and 43.3% part of the time, with the time spent in the class varying between 1% and 99% (OSF, 2019). A total of 26.9% of Tier 3 students studied fully in special education classes in mainstream schools and 8.6% in special classes in special schools (OSF, 2019). The proportion of Tier 3 students assigned to special schools follows a declining trend. The number of special schools has indeed decreased in Finland during the

last decade and thus, the number of students being educated in special schools has halved (Lintuvuori, 2019). Other than that, the trends in Tier 3 students' placements have mainly been quite stable since the three-tiered support model was put into action in 2011 (Figure 2). The small changes in the ratios can be mainly explained by the classification changes in statistical compilation and by the increasing proportion of students defined as students with SEN (Lintuvuori, 2019). Naturally, as the number of students defined as having special educational need continues to grow, the rates for segregated or integrated students raise as well (Richardson & Powell, 2011).

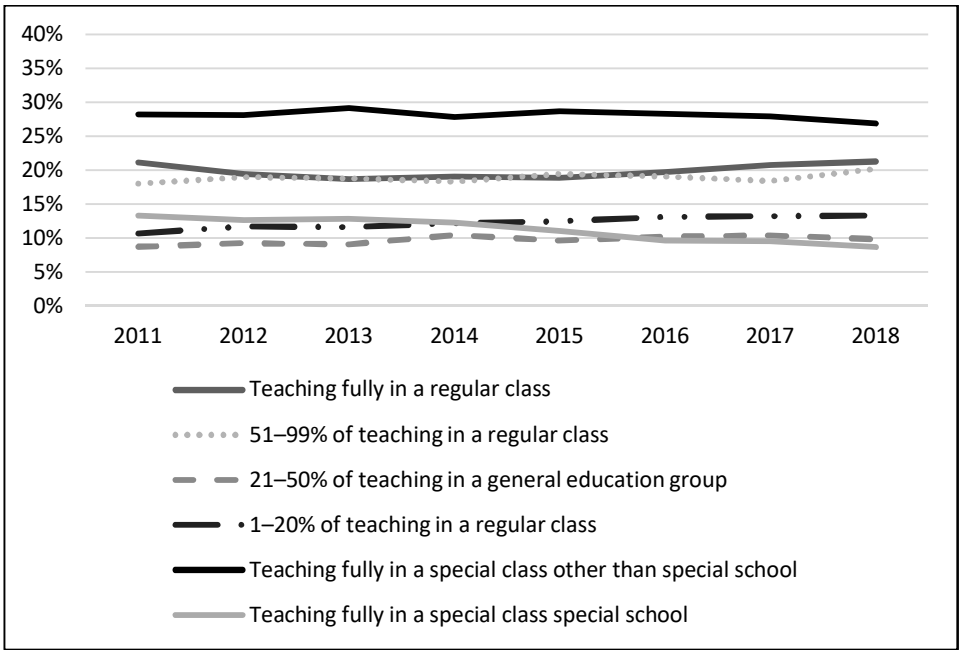

Figure 2 The placement of Tier 3 students in years 2011–2018 (OSF, 2019)

If these statistical proportions are set to the international perspective (e.g. European Commission, 2019), around 4% of all comprehensive school students in Finland study in segregated settings (Lintuvuori, 2019). According to the definition by the European Agency for Special Needs and Inclusive Education (EASIE), segregated setting is defined as follows: "A pupil with SEN follows education in a separate special class or special school for the largest par—80% or more—of their time", whereas the operational definition of an inclusive setting encompasses all education in which "the pupil with SEN follows education in mainstream classes alongside their peers for the largest part—80 per cent or more—of the school week" (EASIE, 2016, p. 9). Yet, as the international classifications differ from the national statistical categories, the indicators for inclusion or segregation as such are difficult to specify.

Like all the averages, the national statistical averages mask variations and hide the fact that there are notable differences across regions and municipalities. For example, the proportion of Tier 3 students studying full-time in regular class can vary from 0% to 100% between municipalities, and when we look at the whole student population, the proportion of all comprehensive school students placed in special classes ranged from 0% to 9% across municipalities (Lintuvuori, 2019).

We can take a closer look at the placement of students with SEN in the present study by exploring the lower secondary education data used in Studies II and III. It was collected in 14 municipalities in the Helsinki metropolitan area and the data used in the analyses consisted of 369 classes. In 45% of regular classes there was at least one Tier 2 student, and in 31% there was at least one Tier 3 student. Overall, there were students with SEN in 58% of regular classes. The average proportion of students with SEN in these classes was 13%, ranging up to 50% (Study II). Other nationally representative data from the lower secondary education has indicated, that there is at least one Tier 2 or Tier 3 student in around 65–80% of regular classes (Hienonen & Lintuvuori, 2019; Kupiainen & Hienonen, 2016) and in primary education, in 80% of regular classes (Kupiainen & Hienonen, 2016).

When dealing with the placement of Tier 3 students, the term *inclusion* is hard to avoid. Broadly, it refers to a process that helps to overcome barriers limiting the presence, participation and achievement of learners (UNESCO, 2017). With the ratification of the UN Convention on the Rights of Persons with Disabilities (United Nations, 2006), several countries have committed to implementing a more inclusive school system and to increasing the proportion of students with SEN in regular schools (European Union, 2018; United Nations, 2006; World Bank 2019). However, the inclusion concept has deliberately been excluded in this study. There were three main reasons for this.

First, in the Finnish basic education legislation, the term inclusion is not used. In the National Core Curriculum, inclusion is mentioned once along with the mission of basic education: “The development of basic education is guided by the inclusion principle” (FNBE; 2016, p. 17). Yet, the term is not explained or defined. Thus, it remains open to various interoperations both in ideological and political senses (Honkasilta et al., 2019).

Secondly, both data sets used in this study include information on student placement (e.g., if a student is placed in a regular or special class) but not the information on how inclusive these regular class placement practices are in reality. As Huber et al. (2001) have noted, inclusive practices should be separated from the inclusion itself—when studying the effects of inclusion, it may be necessary to separate the effects of inclusion itself from the effects of inclusive practices implemented both at a class level and school-wide. Consequently, studying inclusion was not the purpose of the original data collections, thus the data do not contain information on the school- or class-level inclusion policies or practices.

Furthermore, the inclusion cannot be reduced to a mere description of the physical placement; and thus, in this study, instead of inclusion, the terms class placement and class assignment are used (see also Zweers, Tick, Bijstra, & van de Schoot, 2019).

Thirdly, inclusion is an ambiguous term with various definitions (Florian, 2014; Richardson & Powell, 2011). In previous research it has referred to a situation in which students with SEN are taught in mainstream schools and in which students spend at least some of their time in the classroom together with their peers without SEN (e.g., Farrell, Dyson, Polat, Hutcheson, & Gallannaugh, 2007; Huber et al., 2011). Nevertheless, there is no agreement about whether inclusive education refers to both partial and full inclusion (Kauffman, Nelson et al., 2017). There are many ways to provide inclusive education, and there are several reasons for it, e.g., the length of the experience in the implementation of inclusive education, the consistency of educational policy promoting inclusion, and the way inclusion is understood (Szumski, Smogorzewska, & Karwowski, 2017). It is important to note that the educational context in this study only refers to the class the student is assigned to. The idea of inclusive education embraces the heterogeneity of all learners and heterogeneity of all classes (Kiuppis & Sarromaa Haustätter, 2014). Consequently, the terms students with SEN and regular and special classes are not in line with the idea of inclusion. However, in this study, they have been used to make the text more readable.

2.4 Student performance

Although the classroom may have an effect on many things, both on students' cognitive and non-cognitive outcomes, the focus in this study is mainly on the cognitive outcomes. Student- and class-level effects are studied in terms of students' cognitive competencies (Studies I–III), school achievement (Study III), and learning motivation (Study III). This implies an assumption that student's performance is a measurable, objective reality (e.g., Kauffman, Nelson et al., 2017). Hattie (2005) argues that the concept of 'student learning' is usually assumed but not necessarily agreed on. For some it means achievement in curricular domains (such as reading, mathematics), for others it means retaining interest in learning whatever the subject, or task behavior regardless of any changes in test scores. In this study, the effects on students' cognitive competencies are studied in the Finnish learning to learn (LTL) framework and the measure of 'performance' entails students' test scores in several LTL tasks.

There are two main reasons for the use of LTL scores as a measure of performance. First, Finnish comprehensive schools do not administer national, standardized high-stakes tests (Vainikainen et al., 2017). Instead, national sample-based assessments in core subjects, on both regular and irregular bases, have been introduced and established by the Finnish National Agency for Education

(EDUFI), the former Finnish National Board of Education (FNBE). In addition, FNBE began to look for ways to assess more equivocal educational outcomes, i.e. cross-curricular competencies. In the mid-1990s, a group of researchers at the University of Helsinki (later the Centre for Educational Assessment, CEA) started to develop a framework for LTL and tools for its assessment as a task from the FNBE (Hautamäki et al., 2002). Since then, the LTL assessments have been used as one indicator for monitoring the effectiveness of education in Finland. Secondly, as the LTL assessment studies are conducted within CEA, they provided the best possible data for this this dissertation.

One challenge when talking about student performance in general is what has been measured and what terms should be used when referring to the measured outcomes. Learning to learn is an ambiguous term and always bound to the context. In this study, it is defined in the Finnish context (Hautamäki et al., 2002; Hautamäki & Kupiainen, 2014). It may be unfamiliar especially to international readers; and thus, needs more elaboration. However, as the number of words in the original research articles has been limited, it was not possible to present an in-depth introduction of the framework. The concept of learning to learn has been introduced only very briefly in the original publications. Consequently, rather than learning to learn, in all three articles, slightly different terms have been used. Admittedly, as a concept, LTL is open to semantic interpretation (Vainikainen & Hautamäki, in press; see also, Hoskins & Fredriksson, 2008) and the interpretations vary across countries (Crick, Stringher, & Ren, 2014). The decisions on the chosen terms were made according to the editors' and reviewers' comments and suggestions, and in line with the journals' terminology. For this reason, the terms vary between the Studies I–III. In Study I, learning to learn tasks are referred to as *thinking skills*, whereas in Study II, the term *cross-curricular competence* has been used. In Study III, in addition to cognitive learning to learn items, curricular Finnish and mathematics tasks, as well as learning motivation scales were used. Furthermore, as a measure of school achievement, student grades were used. Thus, the terms in Study III are *cognitive outcomes* and *learning motivation*. In this overview part of the thesis, the term *student performance* has been used when referring to the student achievement in LTL tasks. Next, the Finnish LTL framework is introduced briefly. Then, some of the aspects related the measurement of LTL tasks are discussed. The items measuring LTL skills and learning motivation are described in more detail in Chapter 3.3.1.

Learning to learn—ability to adapt to novel tasks

Learning to learn was identified as one of the key competencies by the European Commission in 2006, in the European Framework for Key Competencies for Lifelong Learning. Consequently, as an educational goal, LTL is an explicit part of the EU definition of key competence and of the 21st century skills in the global

context (Vainikainen, Hautamäki, Hotulainen, & Kupiainen, 2015). This has led to the common agreement that in addition to subject matter-specific knowledge, education should enhance more general skills needed in all learning. In short, learning to learn can be seen as an educational outcome that does not fall into the domain of any particular key subject, rather, it can be seen developing in through education in different subjects and thus, seen a common pedagogical goal of all school subjects (Hautamäki & Kupiainen, 2014). Furthermore, LTL skills can be seen both as a process and as an outcome, a measurable one (Crick, Ren, & Stringher, 2014). It is also argued that learning to learn is not only a scholarly concept but it also involves politics (Stringher, 2014). Moreover, there is no agreement in the international literature and in the academic community concerning the concept of learning to learn. The recent National Core Curriculum introduced the concept of transversal competencies and thinking and learning to learn was defined one of the seven competencies. The general idea is that LTL skills are enhanced in all study situations (FNBE, 2016, p. 166). However, as argued, the definitions of LTL differ depending on the context, and in this study, the concept of LTL is understood and defined in the Finnish context following the work of Hautamäki and his colleagues (Hautamäki et al., 2002).

Learning to learn can be conceived as "an ability and willingness to adapt to novel tasks, or the adaptive and voluntary mastery of learning action" (Hautamäki & Kupiainen, 2014, p. 170). It comprises of the initial task acceptance and a learning action that requires and is maintained through both affective and cognitive self-regulation; and thus, it consists of the cognitive competencies and self- and context-related beliefs. Cognitive competence refers both to the knowledge of relevant facts and to the use of thinking and reasoning: how a student can access something that has been learned previously and how a student can apply general, earlier acquired procedures to adapt to the new task (Hautamäki et al., 2002; Hautamäki & Kupiainen, 2014). Beliefs refer to the anticipated emotions which, once activated, lead to either commitment or refusal (Vainikainen et al., 2015). LTL competencies are related to intelligence, understood in a Piagetian framework as the active use of formal operational schemas (Hautamäki et al., 2002; Vainikainen et al., 2015). The orientation of LTL encompasses several theoretical traditions within educational and developmental psychology (Hautamäki & Kupiainen, 2014). In comparison to psychometrically-oriented intelligence research, the major difference is that LTL focuses on the relative malleability of cognitive abilities through educational means (Vainikainen & Hautamäki, in press) and the involvement of motivational and situational factors that are observed as performance scores on given tasks. Neither intelligence nor LTL can be measured as direct observations. It can be seen that LTL tasks measure investments. In other words, solving a task requires investment of inborn abilities, cultivated knowledge and skills that have accumulated during educative processes, and motivation to accept the task and to

try one's best to solve the task (Vainikainen & Hautamäki, in press).In short, in LTL assessment, students are given novel tasks and they are invited to accept them as their own with all the motivational, goal and aptitude related conditions attached to the situation (Hautamäki & Kupiainen, 2014). LTL tasks cover different domains of thinking. According to Demetriou (2014), the spatial, verbal, quantitative, categorical, causal, and, social reasoning systems have been identified by methods from different theoretical origins, and they are accounted as autonomous domains of understanding, thinking, and problem solving. Following the work of Demetriou, Vainikainen et al. (2015) have applied this grouping to the Finnish LTL tasks. Tasks used in this study cover spatial, verbal, quantitative, and causal reasoning and thinking.

In addition to the cognitive tasks, LTL assessment includes the affective dimensions, i.e., both self-related and context-related beliefs and attitudes, assessed by self-reported scales. In the present study, only the goal orientation (Niemivirta, 2004) scales have been used, as a measure for learning motivation (Dweck, 1986). Central to LTL is the interplay of cognitive and affective dimensions, combining both the mastery of thinking and willingness to adapt to a given task.

The tasks may deviate from the tasks those students are used to in ordinary school work, partly because they are not directly derived from any curricular subject and partly because they are modified to include an element of surprise. This means that a student can refuse to attempt to solve the task. A student can also accept the task and attempt it either by solving or not solving it (Hautamäki & Kupiainen, 2014).

The Finnish LTL framework embraces the idea that good quality teaching can enhance both the LTL skills and learning-related attitudes. To some extent, the assessed skills are related to the curricular content, but they require the application of both effort and higher-order thinking skills instead of the repetition of subject-specific knowledge (Vainikainen et al., 2015). In other words, they give an impression of being related to other schools tasks, and at the same time, they require a wider use of skills acquired in the school but also, in learning outside the school (Hautamäki & Kupiainen, 2014).

The cognitive tasks contain 4–16 items, and each affective scale three items. Due to the limited testing time, the scales cover only the most critical areas (Hautamäki & Kupiainen, 2014). All the items and scales have been developed, tested, and redefined since 1996 (Hautamäki et al., 2002). Some of the cognitive scales are modifications of instruments developed by others whereas some have been constructed specially for the Finnish context (Hautamäki & Kupiainen, 2014).

As mentioned, in addition to the cognitive scales, one set of affective scales was used in Study III: achievement goal orientation. Achievement goal theory can be seen as one of the sub-fields in learning motivation theory (Elliot & Dweck,

1988). Generally, four learning goals are distinguished. Mastery orientation entails striving for learning goals whereas performance orientation refers to a student's performance in comparison to others (Elliot & Dweck, 1988). Performance-avoidance orientation reflects the avoidance of demonstrating normative incompetence. Finally, avoidance orientation refers to students' desires to avoid achievement situations and to minimize the effort and time spent on studying. The Finnish LTL framework consists of five goal-orientation scales as the mastery orientation was divided into intrinsic and extrinsic scales (Tapola & Niemivirta, 2008).

As LTL tasks are non-curricular in nature, there can always be an argument about the extent to which they measure school achievement or school performance. However, because they are not bound to any specific school subject, they may assess broader skills and abilities students need when they solve new tasks. Therefore, they can better anticipate students' behavior in future learning tasks—indicating students' aptitude and willingness for learning and self-development (Hautamäki & Kupiainen, 2014). Furthermore, unlike in many standardized tests, there is no effect of teaching to test. Of course, the performance in LTL tasks also reflects the testing situation. It is understandable that not all students enjoy being tested. Furthermore, the tasks are low-stakes tests which entail that an element of motivation is involved in the test situation. However, as in all testing, the idea is the predictive validity and that they predict later learning situations. Furthermore, in the present study, as the focus is on students with SEN, they can be justified. According to Zigmond and Kloo (2017), at its core, special education is about promoting higher order reasoning and problem solving, facilitating independent and cooperative work, and supporting acquisition of new skills and subject matter through both direct and focused instruction. These are also seen in the Finnish context, set as the requirements for the work of special education teachers (Kivinen, 2009).

Some critical aspects of measuring performance

A few aspects related to the LTL assessment and student assessment in general that are worth mentioning here. The first one is related to students with SEN, the second one to the nature of the low-stakes assessment and the third to the test mode.

Various studies have shown that students with SEN, on average perform less-well in tests assessing cognitive competencies (Pohl, Südkamp, Hardt, Carstensen, & Weinert, 2016; Vainikainen, 2014). However, there are always some students with SEN who perform at the same level as students without SEN (Leino et al., 2019; Lintuvuori et al., 2019). Students with SEN form a heterogeneous population with various competence profiles. For this reason, in many countries, larger groups of students with SEN have rarely been included in the national or international large-scale assessments (Heydrich, Weinert, Nusser, Artelt,

Carstensen, 2013). Contrary to many other countries and many other assessment studies, in principal all students are included in the Finnish LTL assessment studies. That is, no student is excluded from the sample based on their SEN status. The decision on whether a student (especially those receiving Tier 3 support) participates in the assessment is always left for the teachers to consider. There are some special classes in both data sets used in this study, for which the teacher estimated, based on prior experience, the assessment tasks to be too demanding. Naturally, since students with SEN cover a heterogeneous group of needs, different challenges are posed when they are included in a large-scale assessment study (e.g., Heydrich et al., 2013). In general, there can be little accommodation for students with SEN (Elliott, Thurlow, & Ysseldyke, 1996), like an Une Heure (UH) test in PISA—a shortened test with easier items (OECD, 2017). In both assessment studies utilized here, all the students completed the same tasks without any modified version for different groups of students. Yet, in retrospect, there are some characteristics that may have increased the test validity for students with SEN.

Firstly, although the instructions for teachers gave the approximation for the test time (4x45min at fourth grade, and 90min at sixth, seventh and ninth grades), the time for a specific task or item was not fixed for students. In general, the time invested in the task is related to better test performance (Kupiainen, Vainikainen, Marjanen, & Hautamäki, 2014), and especially students with SEN tend to benefit from extended time (Sireci, Scarpati, & Li, 2005). However, there are indications that Finnish students with SEN spent less time on task when compared to their peers (Vainikainen & Hautamäki, 2018). Secondly, tests drawing on curriculum-based knowledge may be more challenging for students with SEN due to possible individualized or more restricted curriculum (Heydrich et al., 2013). As the LTL tasks are not directly tied to specific subjects (Hautamäki et al., 2002), students with SEN can also be included better. Thirdly, for students with SEN, it is better if tests are presented in a closed response format (multiple choice) which also allows for an easy and more objective scoring (Heydrich et al., 2013). All the cognitive tasks and affective scales in LTL assessments are measured with multiple-choice questions. However, it has to be noted that students with SEN, on average, might be more prone to engaging in random guessing or even to omitting these items rather than responding by trying to solve complex cognitive items (Gnambs, & Nusser, 2019; Pohl et al., 2016).

Worth mentioning is also the sensitivity of the tasks, especially considering Tier 3 students who are one of the main populations of this study. The LTL assessment is not intended for prognosis at the level of individual students, but rather for class- and school-level diagnostics for educational assessment (Hautamäki et al., 2002). Thus, it has proven its effectiveness at the class-level (e.g., Vainikainen et al., 2015).

The second aspect is that LTL assessments are low-stakes tests. The general expectation is that in high-stakes tests when the test results have important personal consequences for the students, they will put more effort into the test. In low-stakes tests, the stakes are lower at the personal level. Students are expected to balance test taking with other interests, leading to reduced effort, that is, students are not necessarily putting in their best effort (Kupiainen et al., 2014). A related question is whether students with SEN are more or less affected by the low-stakes nature of the tests.

The third assessment-related factor is the test mode. In the primary education assessment, the tests were administered as paper-and-pencil versions in both data collections whereas in lower secondary education, the first measurement was presented in a paper-based booklet and the second measurement was computer-based. A lot of attention was paid to make the two assessments as comparable as possible, yet systematic mode effects cannot be totally ruled out. However, both the previous LTL assessment (Hautamäki, Kupiainen, Marjanen, Vainikainen, & Hotulainen, 2013) and findings from the other mode experiments (Schroeders & Wilhelm, 2010; Williams & McCord, 2006) show that paper-based and computer-based versions are comparable as the differences between them have been negligible.

3 The present study

> "It's very much like you're trying to reach Infinity. You know it's there, but you just don't know where—but just because you can never reach it doesn't mean that it's not worth looking for." (Juster, 2008)

The literature on research and education policy aspects visited in previous chapters has led to the conceptualization and to the research design of this study. In this chapter, I will first introduce the main aims and objectives. Furthermore, I am going to position this study in the field of educational research. After this, data, measures and methodological solutions are described. Then follows a brief overview of the Studies I, II and III.

3.1 Aim and objective

The present study is based on three research articles, which are referred to in the text with Roman numerals (Studies I–III). The general aim of this study is to deepen the understanding of the placement of students with SEN. This aim can be divided into two main lines: first, to discern the class-level effects, specifically, class size and the proportion of students with SEN in regular classes. Second, to explore the effect of the placement by comparing regular and special classes.

To achieve the aims of this study, three objectives are posed as follows:

1) To investigate the effect class size in the Finnish context, more precisely, how it functions as means of support for students with SEN.

2) To study the class composition effect of the proportion of students with SEN taking into account both the effect on the students with and without SEN.

3) To explore the effect of placement focusing on Tier 3 students by comparing the placement in regular and special classes.

3.2 Conceptualization of the study

Research without a context will have no impact, thus, the conceptualization and the design of a study must be presented clearly and critically (Gersten, Fuchs, Compton, Coyne, Greenwood, & Innocenti, 2005). The prerequisite for the

present study is to acknowledge the complex nature of the investigated phenomenon (see, Berliner, 2002).

The main context of this study is the Finnish basic education system. In Studies I–III, the main emphasis is on students with SEN, hence, the sub context of the current study is the multi-tiered support model within the basic education, in other words, special education system. As a field, special education is situated largely but not entirely within general education (Richardson & Powell, 2011). However, it is acknowledged that special education can be more than just a subsystem in general education as it is contextualized as a nested system, not only within general education but also within itself (Richardson & Powell, 2011).

In Studies I and II, all students in the classrooms were included in the analyses. Only in Study III was the focus solely on Tier 3 students. However, as the student allocation decisions, class compositional and placement effects concern all students (Honkasilta et al., 2019; Kauffman, Nelson et al., 2017), this study is located both in the field of educational and special educational research.

In this study, the concept of performance is contextualized in the classroom. Consequently, this study can be positioned in educational effectiveness research (Reynolds et al., 2014). In educational effectiveness research, there is more research focusing on the school effects than on the classroom effects. This is because from an international perspective, there are notable school-level differences and school level is more often representative in international comparisons unlike class level (e.g., PISA). However, many characteristics of the school effect studies can be applied to the class effect studies. Since the Coleman report (1966), it has been understood that the school effects can be a function of a student ability level or prior achievement, meaning that the relationship between prior achievement and later achievement is smaller in some schools than in others. The same idea also applies to the class level. Given the composition of the class, it might be expected that a student in one class will make more progress than the same student placed in another class, depending on the composition of the student body (Blatchford, Goldstein et al., 2002).

Berliner (2002) has proposed that science in education is not a hard science but it is the *hardest-to-do* science. The educational researcher faces varied and heterogeneous participants and, a diverse audience—teaching professionals, parents, policymakers, and a range of stakeholders. Furthermore, education is a question that everyone has opinion on, as we have all been students, and perhaps parents of a student, at some points of our life. Therefore, it is of vital importance to understand that the results of this dissertation do not represent researcher's own opinions, and the aim is not to take a stand or to present the researcher's own preferences. The results of this study are outcomes from the statistical models and analyses that were conducted using the data and previous research. However, the conclusions of this study are a combination of the previous research findings, and

the researcher's deduction; and thus, they reflect the researcher's previous knowledge, thoughts and values.

Quoting Berliner (2002) further, special education research, because of its complexity, may be the *hardest of the hardest-to-do science*. Firstly, one feature the of special education research which makes it more complex is the heterogeneity of the participants (Gersten et al., 2005; Odom et al., 2005). Hence, the challenge is to find populations that are sufficiently homogeneous to form a group and yet large enough to enable statistical analyses (Gersten, Baker, & Lloyd, 2000).

The second feature of this complexity is the educational context. Special education extends beyond the traditional conceptualization of schooling for the majority of students (Odom et al., 2005). The continuum of student placement is broader than of general education (Kauffman, Nelson et al., 2017). Students with SEN can be placed in special classes or special schools, in regular classes, or as a combination of special and regular classes with different proportions of time spent in the classes. Thirdly, it must be acknowledged that students with SEN are often clustered in certain classrooms (Odom et al., 2005). Fourthly, the line that divides students into students with SEN and into other students can sometimes be thin and it can be based on somewhat arbitrary lines (Kauffman, & Lloyd, 2017). Furthermore, special education is always bound in the local context. The placement continuum and the clustering of students are taken into account, as much as possible, when the methodological choices of this study have been done. The definition of students with SEN is discussed in Chapter 2.1 and the variability in the target group has been reflected upon in Chapter 4.2.

The general concept in educational research is that individuals interact with their social contexts (e.g. schools, classes), and therefore the individual students are influenced by the social contexts which they belong to, and the characteristics of the context is influenced by the individuals (Hox, 2010). In general, the individuals and the social groups are conceptualized as a hierarchical system of individuals and groups, with individuals and groups defined at separate levels of this system. Consequently, the complexity of the context must be considered (Berliner, 2002). School can be observed at several levels, and as a result, data with variables observed at distinct hierarchical levels can be produced. In the present study, this leads to research questions that focus on the interaction of variables that describe the students and variables that describe the classes. This line of research requires advanced statistical models.

With this in mind, it is evident that this study operates in the world of quantitative research. The class placement and the class composition effects that follow from the student allocation are modeled and tested statistically. Basically, research questions in education can be grouped into three types: a) description (what is happening?); b) cause (is there a systematic effect?); and c) process or mechanism (why or how is it happening?) (e.g., Odom et al., 2005). The present

study combines the first two questions and hopefully offers future directions for the last one. Odom and his colleagues (2005), in their article on research quality indicators, have stressed the importance of the appropriate match between the question and the methodology. Research questions define the data and the methods. There must be a clear relationship between the data, analyses and proposed questions (Gersten et al., 2005). A key issue is to ensure that analyses and research questions are aligned with the appropriate unit of analysis for a given research question, that is, researchers should define which unit is used in the statistical analyses (Gersten et al., 2005). In the present study, the class is the main unit of analyses (Studies I and II), however, also the student level is considered in all Studies I–III. This study combines correlational and quasi-experimental research designs (see, Odom et al., 2005). The statistical methods of this study have been selected after careful consideration and in order to take into account the two main features of the class level: the nested and hierarchical nature of the research phenomenon (Chapter 3.5.1) and the non-random allocation of students in the classrooms (Chapter 3.5.2).

The development of educational effectiveness research has involved a commitment to advanced forms of statistical analyses that permit the establishment of the relationships between educational factors and student outcomes: structural equation modelling (SEM) from 1970 and multilevel models form late 1980s (Opdenakker & Van Damme, 2000; Reynolds et al., 2014). Following this line of a research, the latter of methodological advances have been used in this study. Methodological solutions and statistical aspects are described in more detail in Chapters 3.5 and 4.3. In this chapter, more general philosophical assumptions are discussed briefly. In Studies I and II, two-level regression models were constructed in order to model the target phenomenon (e.g., Hox & Berghen, 1998; Tarka, 2017). The statistical aspects mostly refer to specific technical-methodological requirements, whereas the philosophical aspects mainly refer to the ontological nature of causality and to the role of modeling in the epistemology of causal inference on the basis of experimental or non-experimental data (Tarka, 2017).

In a philosophical sense, multilevel modeling can be seen to be based on critical realism (Virtanen, Haverinen, & Leskinen, 2018). This entails there being a real world that does not depend on our minds, and that it can be observed; and thus, studied by collecting data on the phenomenon and modeling it. However, the model constructed is based on the data depends on our mind. The conceptual underpinning of the statistical modeling are divided into four parts, following the work of Virtanen and his colleagues (2018) and it is depicted in Figure 3. The four parts are: the phenomenon, theory based on previous research, data, and the statistical model.

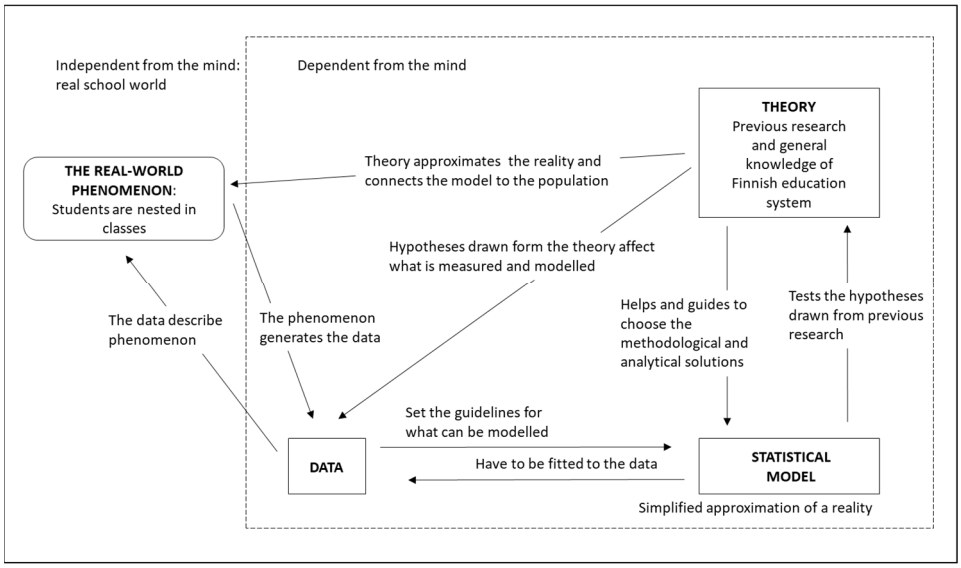

Figure 3 Conceptual model for statistical modeling (Adapted from Virtanen, Haverinen, & Leskinen, 2018)

The *phenomenon* (displayed outside the large box with dashed lines in Figure 3) in the present study entails the students and their nested structure in the classrooms. In general, the complexity of social reality, for example, the latent character of many social phenomena has to be acknowledged. In this study it means admitting that not all the differences between the classrooms can be explained with the modeling or that not all the aspects that affect the between-class differences can be taken into account. Furthermore, some ontological arguments claim that it is not possible to transform the investigated phenomenon into a limited mathematical model (Tarka, 2017). However, Kauffman and Lloyd (2017) argue that even though every important educational question cannot be fully understood as a mathematical equation, many important questions can have mathematical-statistical foundations.

The idea is that the phenomenon is located in the population, and the real-life phenomenon produces the *data* (Virtanen et al., 2018). However, the phenomenon is not completely equivalent to the data. For technical reasons, when we measure something, that measurement will have an error associated with it (Kauffman & Lloyd, 2017; Tarka, 2017; Virtanen et al., 2018). Some of the statistical modeling is to accept that measurement always produces a statistical distribution of values, including errors, and also to understand that error in measurement does not mean that there have been mistakes. It simply means unexplained variation, differences in measurement that have not been attributed to known factors (Kauffman & Lloyd, 2017), which is always considered in the modeling. That is why the *models* have to be fitted to the data and accepted that models represent only approximate estimates of reality (Hox & Becher, 1998; McDonald & Ho, 2002; Tarka, 2017).

There is always a trade-off between the fit of the model and the simplicity of the model. Every model needs a strong theoretical background; in other words, a sound model is always based on the *theory*, which on the one hand is based on findings in the literature, and on the other hand, knowledge in the field, or on researcher's educated guesses. The hypotheses, further operationalization, and the preliminary ideas on causes and effects of variables are drawn from previous research and understanding of the phenomenon being examined (Lei & Wu, 2007). From that follows the specification of the model. If the hypothesized model was only empirically determined, the model would express only the statistical relationships between variables in reference to the analyzed set of data without any reference to actual phenomenon. When the tested model is theoretically determined then it can represent stable causal relationships (Tarka, 2017). Thus, the aim is not only to verify a given theory, but also to conduct this verification on the basis of the measures being analyzed. That is, combination of the theory and the empirical data offers an opportunity for a scholarly explanation of the phenomenon alongside the empirical relationships (Tarka, 2017). The aim in the modeling is two-fold: the model should explain the phenomenon as simply as being possible, even in light of a given theory; whereas the model itself should be coherent with the empirical observations (Tarka, 2017). Finally, it must be accepted that the statistical models are simplified approximations of the reality, not hypotheses that might possibly be true (McDonald & Ho, 2002).

3.3 Measures

Prior to walking the reader through analyses applied in this study, I will summarize characteristics of the data and the measures as they may help to explain why I chose the analytic approaches described later.

3.3.1 Cognitive tasks and motivational scales

Student performance in Studies I–III was measured within the Finnish learning to learn (LTL) framework described in more detail in Chapter 2.4. The LTL scales include verbal, quantitative, spatial and causal reasoning tasks (Vainikainen et al., 2015). The lower secondary education data also contained tasks for curriculum-based Finnish and Mathematics (Study III). In addition to the cognitive outcomes, in Study III, the learning motivation was assessed through five motivational scales taken from the motivational—affective battery of the LTL. Table 1 summarizes all the measures for the cognitive tasks and learning motivation and the number of items.

Table 1 An overview of the learning outcome measures and learning motivation scales

Study	Task	Number of items	Grades
Study I	**Verbal tasks**		
	Hierarchy-rating task	16	4th and 6th
	Short reading comprehension	4	4th and 6th
	Quantitative tasks		
	Mental Arithmetic task	5	4th and 6th
	Hidden arithmetical Operations	4	4th and 6th
	Spatial tasks		
	Piagetian Water-level task	8	4th and 6th
	Dutch geometric analogies	8	4th
Study II	**Verbal task**		
	Missing Premises	7	7th and 9th
	Quantitative tasks		
	Invented arithmetic operations	7	7th and 9th
	Causal reasoning task		
	Control of Variables	8	7th and 9th
Study III	**Verbal task**		
	Missing Premises	7	7th and 9th
	Quantitative tasks		
	Invented arithmetic operations	7	7th and 9th
	Causal reasoning task		
	Control of Variables	8	7th and 9th
	Curricular tasks		
	Finnish	18	7th and 9th
	Mathematics	15	7th and 9th
	Goal Orientation		
	Mastery-intrinsic orientation	3	7th and 9th
	Mastery-extrinsic orientation	3	7th and 9th
	Performance-approach orientation	3	9th
	Performance-avoidance orientation	3	9th
	Avoidance orientation	3	7th and 9th

Cognitive tasks

Next, all the used LTL cognitive task and affective scales are presented by domain, in the same order as the original studies. The assessment in the fourth and sixth grades included more tasks than those used in the present study and only the tasks used in the analyses are presented here. In addition to the cross-curricular LTL tasks, curriculum-based tasks are presented here as well, as they were included in the lower secondary education data.

Learning to learn tasks. In fourth and sixth grade, two verbal tasks were used. The *hierarchy-rating task* was based on Kintch and van Dijk's (1978) model of text comprehension using a Finnish translation of a text regarding the development of US cities in the late nineteen century (Lyytinen & Lehto, 1998). In the task, students were asked to read a one-page text and to assess 16 statements using a 3-point scale to determine whether they presented a good description of the text as a whole, important information regarding the content of the text, or referred to less important details in the text (Hautamäki & Kupiainen, 2014). The second verbal task was a short text set in a context closer to everyday life, which was adapted from a Finnish Vocational Guidance Office test. It assessed students' ability to understand complex sentences and analyze and interpret written information with four multiple-response items (Vainikainen, 2014). In lower secondary education, one verbal task was used. The *Missing Premises task,* adapted from the Ross Test of Higher Cognitive Processes, covered seven items (Ross & Ross, 1979). The students were presented with a fact (premise) and a conclusion, and their task was to choose from among five alternatives the second fact (premise) which would make the conclusion valid (Hautamäki & Kupiainen, 2014).

In fourth and sixth grade, two quantitative tasks were employed. One was the *Mental Arithmetic task*, an adaptation of the Arithmetic subscale of the Wechsler Adult Intelligence Scale - Revised (WAIS-R: Wechsler, 1981). The teacher read aloud eight mathematical problems (e.g. If you buy two bus tickets and one ticket costs 3 euros 50 cents, how much money do you get back if you give 10 euros?), and the students wrote down the answer in their test booklets (Vainikainen, 2014). The other one was the *Hidden Arithmetic Operators task* based on the quantitative-relational arithmetic operator task of Demetriou, Pachaury, Metallidou and Kazi (1996). The task comprised problems with one to four operators (e.g., (10 a 5) b 1 c 2 = 7), and the operations were marked with letters a, b, c and d, standing for different operators: addition, subtraction, multiplication and division (Hautamäki & Kupiainen, 2014). The student's task was to reason which operation the different letters stood for in each case. In the seventh and ninth grade, quantitative reasoning was assessed by seven items based on the *Invented Arithmetic Operators* task which is a modified version of Sternberg's Triarchic Test Creative Number Scale (Sternberg, Castejon, Prieto, Hautamäki, & Grigorenko, 2001). In the task, arithmetic operators were conditionally defined

depending on the value of the digits they connect. The tasks used two invented operators (*lag* and *sev*) and it could comprise several operators in the same equation (e.g., "x lag y = x + y, if x < y, otherwise, x lag y = x – y"; What is 2 sev 3 lag 4?).

Spatial reasoning was covered with two tasks in fourth and sixth grades. It measured by the classical water-level task of Piaget and Inhelder (1956). The students were asked to draw the lines on a picture of eight empty bottles indicating the water level and marking the area filled with water when the bottles were half full. One of the bottles was standing, and the rest of them were tilted on several angles: 45°, 90°, 135°, 320°, 270°, 225° and 180° (Vainikainen, 2014). In addition, the *Analogical reasoning task* was included in the assessment in fourth grade. The task was adapted from a Dutch geometric analogies test (Hosenfeld, van den Boom, & Resing, 1997). The students were presented with eight pairs of geometric figures, e.g. a small square on the left and a big square on the right. The task was to apply the same rule when the student had to choose a pair from five options for another figure (e.g. a small circle). The transformations included adding an element, changing sizes and positions, halving and doubling, and the maximum number of simultaneous transformations was three.

In seventh and ninth grades, the causal reasoning was assessed by the *Control of Variables task*, which is a modified version (Hautamäki, 1984) of the Science Reasoning Task 'Pendulum' (Shayer, 1979) regarding the control of the variables and it is based on one of the schemata identified by Piaget and Inhelder (1958). The students were presented with comparison sets in the world of Formula 1 races with four variables (driver, car, tires, and track) and they had to judge whether the single effect of a variable could be concluded from the comparison (Hautamäki & Kupiainen, 2014).

Curriculum-based tasks. In addition to the cross-curricular LTL tasks, the Metropolitan longitudinal study included tasks for curriculum-based Finnish and mathematics. These specific subjects were chosen because it is generally agreed that skills in these are important for students' future success. The tasks were designed to measure contents and objects students are expected to acquire by the end of the sixth grade determined in the National Core Curriculum (FNBE, 2004) as the first data collection took place already at the beginning of the seventh grade. The aim when designing the tasks was that half the students would be able to solve them correctly and that the scores would follow the normal distribution. Finnish was assessed using18 items and mathematics using 15 items.

As presented in Table 1, in Study I, cognitive tasks from grades four and six were used. The cognitive tasks in the sixth grade were similar to those in the fourth grade (Vainikainen, 2014). However, in the sixth grade, more difficult items replaced the easiest items from the fourth grade. Only items which were identical for both grades were used in the analyses. Five out of the six tasks were included in the model from both grades, the analogical reasoning was included only from

the fourth grade to the analyses to control for the initial differences between students. The items in all tasks were coded dichotomously as correct or incorrect, and a mean for the percentage (0–100%) of correctly solved items was calculated for all 20 items, for both grades. The reliabilities, calculated as Cronbach's alphas, were .75 in the fourth grade and .83 in the sixth grade. For analogical reasoning, items were scored dichotomously as correct or incorrect, and the mean of the percentage of correctly solved items was calculated; the reliability was .77.

In Studies II and III, the LTL items and the curricular Finnish and mathematics items were identical for seventh and ninth grades. The answers were coded dichotomously for a correct answer to all the items in the tasks. For the analyses, in Study II, a mean for the percentage of correctly solved items was calculated based on the 22 items and the reliabilities were acceptable, $\alpha = .84$ in the seventh grade, and $\alpha = .87$ in the ninth grade. In Study III, all the tasks, including the curricular tasks, were treated individually in the analyses. Reliabilities across the sample in seventh grade (Cronbach's α) were .49 for verbal tasks, .75 for quantitative tasks, .79 for causal tasks, .78 for mathematics, and .63 for Finnish. Equivalent figures in the ninth grade were .59, .77, .82, .82, and, .75, respectively.

Goal orientation

In general, the LTL assessment includes self-reported questionnaires for the affective domain, and they comprise scales for several factors relevant to new learning, and to school achievement (Hautamäki & Kupiainen, 2014). In the Study III, the scales for learning motivation were used. In the Finnish LTL framework, an extended five-dimension model for achievement goal orientation was applied (Niemivirta, 2004). Learning motivation was assessed with five goal-orientation scales that students completed during the assessment (Hautamäki et al., 2002): 1) mastery-intrinsic orientation (e.g., "To learn as much as possible is an important goal for me at school," $\alpha = .85$); 2) mastery-extrinsic orientation (e.g., "Getting good marks at school is important to me," $\alpha = .80$); 3) performance-approach orientation (e.g., "I feel I have reached my goal if I do better or get a better mark than most of the other students," $\alpha = .71$); 4) performance-avoidance orientation (e.g., "For me it is important not to fail in front of my classmates," $\alpha = .78$); and 5) avoidance orientation (e.g., "I have no interest in doing anything extra for school," $\alpha = .68$).

The scales offered statements on which the students were asked to take a stance in terms of the degree to which the statement reflects their opinions, their view of themselves, or their mode of action in different situations, for example (Hautamäki et al., 2002). The response scale was a seven-point Likert-type scale with only the end points given a verbal description (1 = not at all and 7 = yes, exactly so). Mean scale scores were computed for each student for the ninth grade.

3.3.2 Background variables

Student characteristics in general can be divided into demographics (e.g., age, gender, socioeconomic status), ability (e.g., performance in LTL tasks), attitudes (e.g., learning motivation), and behaviors (Lee, 2000). Moreover, we may consider these variables measured either on students as statistical controls or as social distribution parameters that we investigate as functions of school characteristics.

Various background variables were used to describe the participants in each study. In addition, they were used as explanatory variables in the regression models (Studies I and II), as covariates to predict the placement and as outcomes of the placement effects (Study III). The descriptive statistics for the variables were presented in the original publications and here, only the overall descriptions on the variables will be provided.

Students' gender and the time of birth (month and year) were extracted from the background information questionnaire presented to the students at the beginning of the assessment. The socioeconomic status (SES) of students was measured as mothers' educational level which was also extracted from the student background questionnaire (Studies II and III). Originally, educational level was asked for with a seven-category question, however, the answers were recoded to three levels: basic education level (only compulsory education); secondary level (upper secondary school or vocational training), and tertiary level (polytechnic or university).

Students' SEN status in the fourth and sixth grades was measured by asking the class teachers to report whether the student had received Tier 2 intensified or Tier 3 special support (Study I). In the lower secondary education, it was measured at the ninth grade by asking special education teachers to complete the questionnaire about information on whether the student received Tier 2 or Tier 3 support. In addition, special education teachers were asked to provide information about whether a student at the Tier 3 level studied according to the general or an individualized curriculum, and to list all the subjects student studied according to the individualized curriculum. For the analyses (Study III), the question was coded according to a national statistics five-category classification (OSF, 2019): 1) general curriculum; 2) individualized in one subject; 3) individualized in two to three subjects; and, 4) individualized in at least four subjects. The fifth category, studying according to functional skill areas was not applicable in the data.

In Study III, each student's school achievement was measured with grades at ninth grade that were derived from the National Joint Application Register. The grades range from 4 (failed) to 10 (excellent). The grades were analyzed separately and, in addition, the grade point average (GPA) was calculated as an average of Finnish, mathematics, foreign language, and science (i.e. an average of geology, physics, chemistry, and biology).

All three studies were undertaken using information on class placement of students. Students' class was determined by the class the student was assigned to in the school register, like 4A or 6B. In all data sets, this placement information was extracted from the school-based lists provided by the education department of each municipality. All classes at each grade participating in the data collection was sampled within each school which made possible to study the differences between classes. Class size was based on student lists, and it was used as an explanatory variable in Study I, and as a one of the criteria for coding the classes as regular and special classes in Studies II and III.

Due to the nested structure of both sets of data, in addition to the student-level variables, class-level variables were used. Some of the class-level variables were aggregated student-level variables. In Studies I and II, the class composition, more precisely, the proportion of students with SEN in class, was calculated by aggregating the support received to the class level as a mean percentage of the Tier 2 and Tier 3 students (0–100%), and in both studies, it was used as an explanatory variable at the class level. In Study III, the proportion of Tier 3 students was calculated in the same way, and it was used as a criterion for coding the classes as regular and special classes. Furthermore, in Study I, the fourth grade test scores and in Study II, the seventh grade test scores (a mean for the percentage of correctly solved items of all tasks) was aggregated to the class level to control for the initial performance differences of students.

3.4 Samples and Participants

The data for this study were drawn from two research projects in order to cover both the primary (Study I) and the lower secondary education (Studies II and III). I participated in all stages of the data collections and hence, became familiar with the data before conducting the present study. As the data are described in more detail in the original studies, only a brief overview is provided here.

3.4.1 Study I

Study I used the data from a nine-year longitudinal learning to learn study conducted in a large municipality in Finland. The study was conducted by the Centre for Educational Assessment at the University of Helsinki on assignment from the Education Department of the City. In autumn 2007, 16 schools were randomly selected from the schools in the municipality using an equal-probability method to ensure representativeness with regard to socio-economic status. One of these schools refused. At the beginning of the fourth year, four new schools were included in the study as some of the original sample students had transferred to them. Finally, the study was extended to include the whole age cohort of the 20 (16+4) original schools which took part in the study at the beginning of the fourth

grade in 2010 and at the end of sixth grade in 2013 (Vainikainen, 2014). The present study was based on data from the fourth and sixth grades. Altogether 978 students attended these classes during the assessments, but only students providing data in both data collections were included in the analyses. In addition, classes with fewer than 10 students were excluded from the analyses as the focus was on regular classes. Thus, the final number of students in the further analyses was 869 (52% girls) from 45 classes and from 20 schools. Of the students with SEN, 61.9% were boys and of students without SEN 46.4% were boys. The mean age of these students at the time of the fourth-grade data collection was 10.23 years (*SD* = .33) and the sixth-grade data collection 12.81 years (*SD* = .33).

3.4.2 Studies II and III

In Studies II and III, the data were drawn from the Educational Outcomes and Health of Children in the Differentiating Helsinki Metropolitan Area study that combines the learning and wellbeing of the lower secondary education students in 14 municipalities of the Helsinki Metropolitan Region, the largest urban area in Finland (Vainikainen & Rimpelä, 2015). It is a Finnish Academy-funded study in which data collection was partly funded by the participating municipalities, and it was conducted by researchers at the University of Helsinki, the University of Tampere, the Finnish National Board of Education, and the National institute for Health and Welfare. The data collecting was carried out by the Centre for Educational Assessment at the University of Helsinki. The first phase of the data collection was conducted in autumn 2011, at the beginning of lower secondary education at seventh grade (*N* = 10 364) and the second phase, with the same students, in spring 2014, at the end of lower secondary education at ninth grade (*N* = 9 441). The data used in this study consist of assessments completed by students at both stages of the study, and for whom the information on SEN status was available.

For Study II, classes with fewer than 11 students were excluded from the analyses as they were mostly special education classes. Furthermore, classes with fewer than 10 students present at the time of the assessment were also excluded from the analyses as their class-level results were not seen as representative. The final number of students in the analyses was 5368 students (49.8% girls) in 359 classes in 96 schools. Of the students with SEN, 60.8% were boys, and of students without SEN, 48.9% were boys. The mean age of students at the time of the ninth-grade data collection was 15.9 years (*SD* = 0.33).

In Study III, the focus was on students at Tier 3 level (*N* = 860, 31.70% girls). The students in special classes were included in the analyses (*N* = 413), also in regular classes (*N* = 447), in cases when there were Tier 3 students in a class. That is, Tier 3 students in special classes, who had been excluded from Study II, were

now included in the analyses. The mean age of students at the time of the second collection was 15.42 years (*SD* = 0.86).

Procedure

Students completed the tasks and the questionnaires in a normal classroom setting as a part of an otherwise regular school day. The teachers administered the assessment by the written instructions. Detailed and standardized test instructions ensured that all participants shared the same information about the tasks. The students were allocated 4x45 minutes at the fourth grade, and 90 minutes at the sixth, seventh and ninth grades for the assessment. The time had proven sufficient in previous assessments (Hautamäki et al., 2002).

In the fourth, sixth, and seventh grade, the tasks and questionnaires were presented in paper-based booklets. In the ninth grade, the entire test was administered on the computer. Previous studies have shown that the paper-based versus computer-based versions are comparable (Hautamäki et al., 2013; see also Schroeders & Wilhelm, 2010; Williams & McCord, 2006).

Every data collection started with the background questionnaires for students. The affective scales were divided into three (primary education) or two (lower secondary education) sets and dispersed between the cognitive tasks to avoid unwanted interaction. For example, the questionnaire for the assessment of learning motivation was presented to the students before any competence tasks were given (Hautamäki et al., 2002).

At the scoring phase, it was ensured that it was consistent and reliable across all data collectors and scorers (Gersten et al., 2005). The cognitive tasks were all in an easy-to-answer-easy-to-score multiple-choice format (Hautamäki et al., 2002). In the paper-based booklets (grades 4, 6 and 7), the researchers coded the items based on the correct answers as correct or incorrect. In the computer-based test (grade 9), the items were coded instant as correct or incorrect.

3.5 Methodological solutions

In this chapter, I will present an overall account of the methodological choices made in this study, along with the theoretical justifications for the methodological solutions made in the original studies. The prerequisites and some of the limitations of the methods are discussed here. A summary of the methodological solutions is given in Table 2. Here, a general rationale of the methodological approaches is provided as the more detailed descriptions of the used analyses and the key descriptive statistics are presented in the original studies. The limitations and strengths of the multilevel models are discussed in more detail in Chapter 4.3.

According to Gersten and his colleagues (2005), a choice of appropriate data-analytic strategy is partly art and partly science. The art is to determine strategies that are powerful enough to detect the investigated effects, while also

sophisticated enough to cover the complexities of the phenomenon. In addition, the methodological strategies have been determined within the limits of the data. This study is based on utilizing existing, secondary data, thus, the selection of the best possible methodological solutions has been a constant compromise between the limits of the data and the requirements of the research questions.

The methodological solutions of this study can be divided into two main lines: in Studies I and II, multilevel modeling techniques were employed, more precisely, two-level regression models were specified to discern the student- and the class-level effects (Table 2). In Study III, the main aim was to investigate differences across the two groups. In order to do that, a quasi-experiment was created using propensity score matching. All the analyses were undertaken using SPSS 24 (Studies I–III) and in MPLUS Version 7.2 (Studies I and II). In Chapters 3.5.1 and 3.5.2, the two main methodological solutions are considered from the following two points of view: 1) the reasons for choosing those methods and, 2) the fundamental nature of the methods.

Table 2 Summary of data and methodology

	Study I	Study II	Study III
Main aims	To test whether class size is positively related to performance at the beginning of fourth grade and, at the end of sixth grade. In addition, to test the assumption that students with SEN would benefit from smaller classes	To investigate how the proportion of students with SEN in class is related students' performance at class level, and to test whether the effect is different for students with SEN than for students without SEN	To study the differences at the beginning of seventh grade and at the end of ninth grade between Tier 3 students assigned to two different educational settings: special and regular classes
Participants	Same students in the fourth and sixth grade (*N* = 869) from 45 classes. All tier levels included.	Same students in the seventh and ninth grade (*N* = 5368) from 359 classes. All tier levels included except Tier 3 students in special classes.	Same students in the seventh and ninth grade (*N* = 860) from 150 classes. Paired data contained 268 students. Only Tier 3 students included.
Focus of the study	Regular classes	Regular classes	Regular and special classes
Measures	Dependent: LTL tests score as the mean for correctly solved items (6th grade). Predictors: LTL tests score as the mean for correctly solved items (4th grade), Analogical reasoning, Class size, % students with SEN in class	Dependent: LTL tests score as the mean for correctly solved items (9th grade). Predictors: LTL tests score as the mean for correctly solved items (7th grade), age, SEN status, % students with SEN in class	Dependent: curricular Finnish and mathematics, Goal orientation scales, GPA and separate grades in the ninth grade Matching variables: Gender, Age, SES, Curriculum, seventh grade LTL tasks
Statistical methods	Two-level regression model	Two-level regression model	Propensity score matching, independent samples t-test, ANOVA
Levels of analysis	Student and class	Student and class	Student
Statistics of interests	Regression coefficients and correlations between variables	Regression coefficients and bivariate correlations. Fixed and random slopes, cross-level interactions	Differences in group means

In general, one of the main challenges in studies investigating the class level effects is that the allocation of students to different classes is not a random process (e.g., Harker & Tymms, 2004; Hoxby, 2000; Konstantopoulos & Traynor, 2014; Paufler & Amrein-Beardsley, 2013). The non-random nature of the processes behind allocating students to classes has been considered in Studies I and II in two ways. Firstly, with longitudinal data, students' prior achievement (i.e. LTL task performance) was used to control for the pre-existing differences between students and classes. In Study III, the propensity score matching method has been employed to create a statistically equivalent experiment and comparison groups and thus, any differences between the groups should have reflected more of the true class placement effect than the initial differences between students.

3.5.1 Multilevel models—students and classes as units of analysis

When we want to explain the different academic achievement of students, prior achievement and different student background characteristics are found to be important factors (e.g., Harker & Tymms, 2004; Thrupp, Lauder, & Robinson, 2002; Wilkinson et al., 2002). However, there remains unexplained class-level variance, that is, variance not explained by the individual differences. This is one of the starting points in this study, to detect the variation that lies between classes. This can be called as compositional effect described in more detail in Chapter 2.2.3.

To analyze the compositional effects such as the class size and classroom composition, and to take into account the hierarchical and nested structure of the data, multilevel methods are the appropriate methodological approach (Bryk & Raudenbush, 1992; Byrne, 2011; Gersten et al., 2005; Heck, Thomas, & Tabata, 2010; Hox, 2010; Lee, 2000). With multilevel techniques, the dependence among the individual responses within the same class can be considered. This dependence may arise because of the shared experiences within the class or the ways the individuals are initially assigned to classrooms (Bryk & Raudenbush, 1992; Heck et al., 2010), or in the case of this study, both. In short, students are nested within classes, and the logic of the compositional effects suggests that the classes should possess similar normative climates, since their members interact closely each day and possibly over a long period of time (Dreeben & Barr, 1988). Thus, we assume that students in the same class are more alike than students from different classes as they are supposedly exposed to the same conditions in their immediate learning environment, whereas students of different classes have different learning environments (e.g., Heck et al., 2010; Scharenberg, 2016).

In general, multilevel modeling allows the data to be presented in their proper hierarchical locations exploring relationships among the variables located at the different levels simultaneously (Heck et al., 2010; Lee, 2000; Lüdtke et al., 2008), in this case, at the student and at the class level. Naturally, classes are also nested

within schools. However, school level was not included in the analyses because it would unlikely add much over the class level as the differences between schools tend to be rather small in Finland (e.g., OECD, 2016b, p. 226).

These hierarchical models involve a search for statistical associations between variables measuring students, on one hand, and class level factors on the other (Lee, 2000). Furthermore, multilevel procedures enable investigation of the variation in outcomes in student performance that exist at different levels of the data hierarchy. It therefore becomes possible to develop more refined conceptual models about how the explanatory variables at each level contribute to the variation in outcomes (Heck et al., 2010). Lastly, multilevel models also permit variability in the regression coefficients (slopes) to be studied. Therefore, whether the strength of a relation between an explanatory variable and a dependent variable vary across classes can be studied (Heck et al., 2010). In addition, cross-level interactions can be defined to indicate interactions across levels of the data hierarchy. To sum up, one of the aims in multilevel models is to test how variables measured at one level affect relationships occurring at another level (Bryk & Raudenbush, 1992).

Multilevel models ensure that some of the challenges occurring with single-level models (i.e., analysis of variance, regression analysis) can be solved, such as aggregation bias, misestimated standard errors, and heterogeneity of regression (Lee, 2000). Firstly, aggregation bias can occur when a variable takes on different meanings at different levels and therefore, has different effects at different levels of aggregation. For example, the proportion of students with SEN in class may influence a student's performance above and beyond student's own SEN status. Aggregation bias in multilevel models can be avoided because it is possible to investigate the effects of a similar phenomenon at more than one level of aggregation simultaneously (Cheung & Lau, 2008; Lee, 2000). Secondly, with multilevel data, the estimation of the standard errors can lead to misestimates if the individual cases are treated as though they are independent even though they are not. For example, students' performance in the same classes shares at least some dependence on other classmates (Lee, 2000). The multilevel models, as described above, take the nested structure of the data into account. A third challenge concerns heterogeneity of the regression slopes. In other words, relations between student characteristics (such as SEN status) and performance (such as LTL test scores) can vary across classes and may be functions of the class level variables (Lee, 2000). These challenges are taken into account in multilevel models.

In classes, there are individual students, their background characteristics, prior knowledge and achievement, and learning aspirations (Micro level) (Figure 4). The students are nested within classes with a given number of students and characteristics that are aggregations of the individual student characteristics (Macro level). Consequently, drawing on a range of theoretical perspectives, in

Studies I and II, two-level regression models were proposed, with students (Micro level) nested within classes (Macro level) to account for the possible dependency effects among students clustered in the same class. The two levels are referred to as Level 1 and Level 2, respectively.

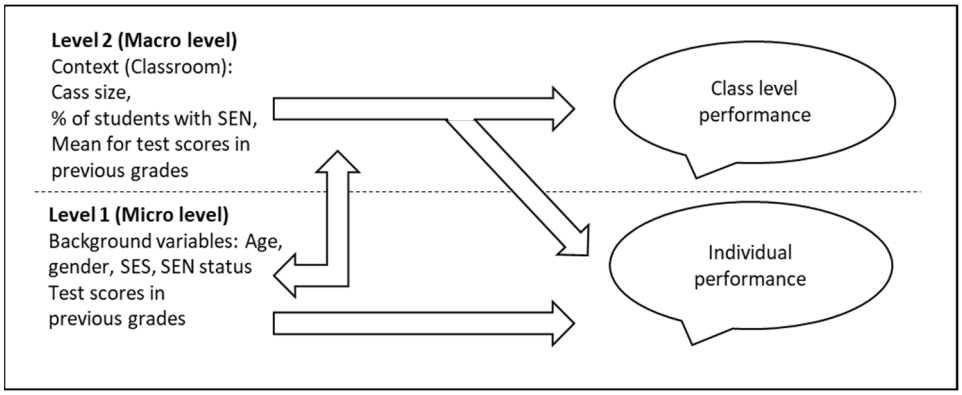

Figure 4 Defining variables in a two-level model (Adapted from Heck et al., 2010)

Compositional effects are often detected with the same variables: these phenomena can be expressed simultaneously in both individual and aggregate forms (two-headed arrow in Figure 4). Student level variables are measured on students. However, the class level variables can be derived from two different sources (Lüdtke et al., 2008). The first can be measured directly (e.g., class size) and these variables that cannot be broken down to the individual level. The second is produced by aggregating variables from a lower level (e.g., average performance level of a class). Both types of variables can be seen as contextual variables. Level 2 variables can be presented both for the purpose of statistical controls and as class characteristics (Lee, 2000). In other words, two types of outcomes as functions of class characteristics can be explored: prior achievement in LTL tasks can be treated as a control for the average performance level of a class (mean performance) and the proportion of students with SEN in a class as a social distribution parameter. After controlling for the average class-level achievement, the effect of the proportion of students with SEN can be explored. As depicted in Figure 4, the models are based on the assumption that there is a potential relationship between the different student background variables and the student level performance, and between the context variables and the class level performance, represented by horizontal arrows. Finally, multilevel procedures also enable study of the effects of the explanatory variables at a higher level of the data hierarchy on a relationship at a lower level (arrow that extends from the macro level towards the micro level).

Following Hox (2010), modeling in this study was done with three distinct steps. In order, they are: 1) specification of the baseline model (model with no predictors); 2) specification of the Level 1 model; and, 3) specification of the Level

2 model (see also Heck et al., 2010). The baseline model for student *i* in class *j* can be presented as follows:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij} \quad [3.1]$$

where β_{0j} is the intercept and ε_{ij} represents the errors in estimating individual test scores within class (*j*). In the equation, the subscript (*i*) is for students and (*j*) for classes. Between classes, variation in intercepts can be presented as:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad [3.2]$$

where γ_{00} refers to the overall intercept across classes. Variation in estimating class level intercepts is represented as u_{0j}.

Via substitution, the baseline model can be written as:

$$Y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij} \quad [3.3]$$

Multilevel models provide the partitioning of the total variance between students (ε_{ij}) and classes (u_{0j}). As described, multilevel models are needed because the nested data violate the assumption of the independence of all observations. This amount of dependence can be expressed as the intraclass correlation (ICC) (Bryk & Raudenbush, 1992; Byrne, 2011; Heck et al., 2010; Hox, 2010; Lee, 2000; Maas & Hox, 2005). The ICC can be defined by the equation:

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e} \quad [3.4]$$

For example, ICC .30 indicates that 30% of the total variability in LTL scores lies between classes. This first step also indicates whether multilevel modeling is needed: when the ICC is greater than 10% of the total variance in the outcome the multilevel methods are needed (Lee, 2010).

To begin with, the modeling started with the decomposition of the variation into within- and between-class components. Hence, the baseline models were defined, where the variance was divided into the student level and to the class level, and no student or class characteristics were considered yet. In both studies, by exploiting the multilevel approach, the class level accounted for close to 20% of the variation in the LTL performance (Study I, 17%, and Study II, 18%). The amount of class-level variance is quite typical in Finland (Kupiainen, 2019; Thuneberg et al., 2015¸ Yang Hansen et al., 2014). Furthermore, in educational

and organizational research, ICC rarely shows values greater than 30% (Lüdtke et al., 2008).

Next, following Heck et al. (2008), in both studies, models were built up by entering one variable at a time. This forward-stepping strategy was used to explore the effects of different variables (Bryk & Raudenbush, 1992).

In its simplest form, an example of a student level model can be expressed as:

$$Y_{ij} = \beta_{0j} + \beta_{1j}SEN_{ij} + \varepsilon_{ij} \quad [3.5]$$

To explore the equation closer, the variable Y*ij* is the outcome for a student *i* in group *j* predicted by the intercept β_{0j} of group *j* and the regression slope β_{1i} in group *j*. To be precise, β_{j1} refers to the slope for the relationship in class *j* (class level) between the student level predictor (student's SEN status) and the dependent variable (LTL test scores). It simply indicates the amount by which the mean test score changes, when the explanatory variable (SEN status) changes one unit (Bryk & Raudenbush, 1992).

The third step was to add all the class level variables in the model, again one by one. This can be defined in its simplest form as follows at a class level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}SEN\%j + u_{0j} \quad [3.6]$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad [3.7]$$

where γ_{00} refers to the overall intercept. This is the mean of the LTL scores on the dependent variable across all the classes when all the predictors are equal to 0. In equations 3.6 and 3.7, γ_{01} refers to the overall slope, between the dependent variable and the class level predictor. Subscript $^{SEN\%j}$ is the class level variable, proportion of student with SEN in class. As mentioned earlier, the effects of the variables in the model (e.g., SEN status) can be seen as deriving from two sources: the variable's effect on students, and its additional effect when aggregated. A multilevel model allows these to be separated, as shown above. Finally, γ_{10} refers to the overall slope, between the dependent variable and the student level predictor.

Model estimations in Mplus were done according to Muthén & Muthén (2010) (see also Hox & Bechger, 1998; Lei & Wu, 2007). Because the variables were close to normally distributed and the sample sizes were sufficient, maximum-likelihood robust (MLR) estimation was used. It is known to be robust to non-normality and to allow for unequal group sizes (Byrne, 2011, p. 349). The goal is to determine whether the hypothesized and estimated models are consistent with the data collected to reflect the hypotheses (Lei & Wu, 2007). The consistency is evaluated through model–data fit, which indicates the extent to which the

postulated relations among variables are plausible. As incremental fit indices, to measure the increase in fit relative to a baseline model (Lei & Wu, 2007), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI) were used in Study I. In Study II, the Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted Bayesian information criterion (aBIC) goodness-of-fit values were used (see, Hox, 2010; Byrne, 2011). In all, both models fit the sample data reasonably well as indicated by the selected overall goodness-of-fit statistics. Furthermore, Root Mean Square Error of Approximation (RMSEA) was used to assess how well the estimated model approximated the true model, and it was within the recommended limits (Hox & Becher, 1998; Kline, 2005).

For readers who are not familiar with the equations, a good way to conceptualize and to describe a model is to present it in a graphical form (Greenland, 2017; Hox & Bechger, 1998; Lei & Wu, 2007). In Figure 2, the overall form of the proposed baseline two-level regression model applied in Studies I and II, is presented. As depicted below, multilevel models allow a simultaneous estimation of effects at the micro- and macro level (also, Figure 4) (Byrne, 2011; Heck et al., 2010). The basic multilevel model treats the Level 1 intercept as an outcome with variance (elliptical shape in Figure 5) that can be explained using variables from a higher level (Heck et al., 2010). In the figure, single-headed arrows represent regression coefficients and they are used to define a relationship in the model (the variable at the tail of the arrow causing the variable at the point) (Hox & Bergher, 1998). Double-headed arrows indicate covariance or correlations between the predictor variables, without any causal interpretations. To be specific, the regression coefficients in the student level for each class are conceived as outcome variables that are considered to depend on specific class level characteristic (Bryk & Raudenbush, 1992).

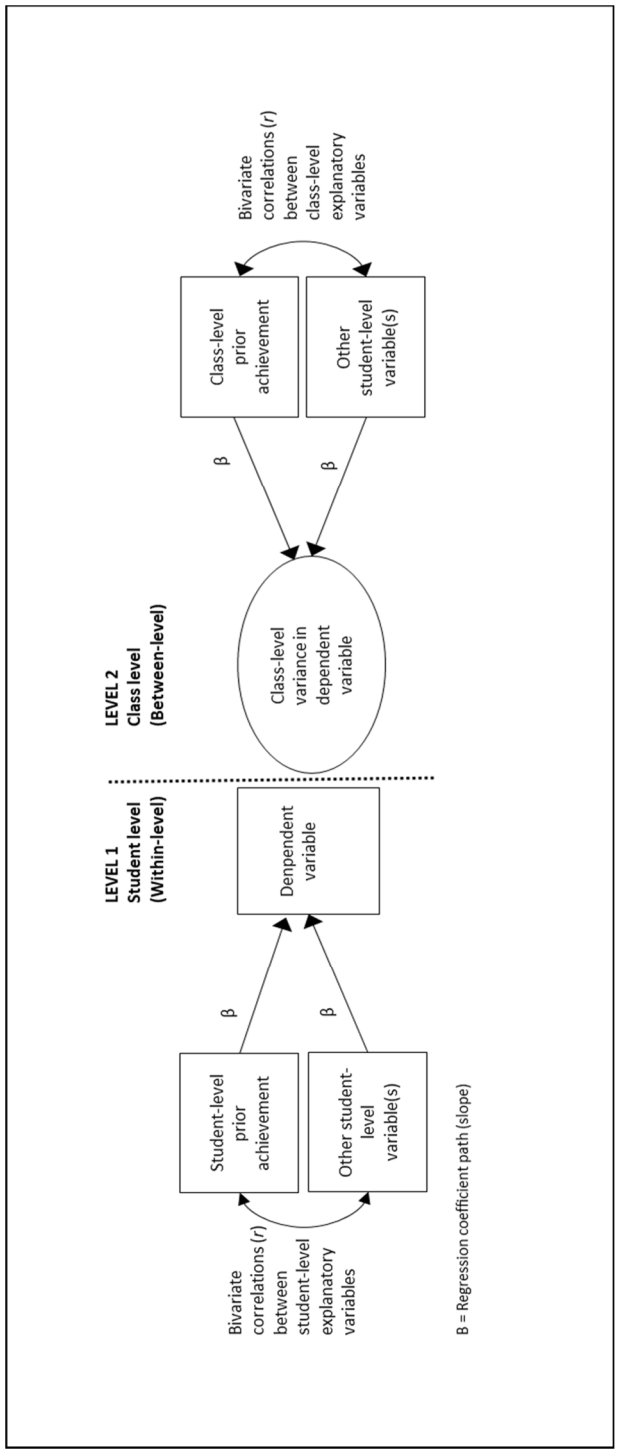

Figure 5 A proposed two-level regression model in Studies I and II

So far, the models have been described in a way that the between-class slopes have been treated as fixed, that is, the effect is fixed to be the same for all classes (Heck et al., 2010). However, one of the hypotheses of Study II proposed that the relationship between student performance and the proportion of students with SEN in class may vary across classes. Therefore, the between-unit slope (i.e. the proportion of students with SEN in class and the class-level performance) was defined to be varying randomly across units. In other words, the slopes were different across classes. Finally, the cross-level interaction was studied. This means that how the class-level explanatory variable might explain the variability in the student-level variable was detected (Heck et al., 2010). Cross-level interaction is depicted in Figure 4 with an arrow that extends from Level 2 towards Level 1 This type of effect implies that the magnitude of a relationship observed within classes is dependent on contextual features defined by higher level units (Heck et al., 2010). Interactions indicate that the relationship between a predictor (SEN status) and the outcome (LTL test scores) depends on the value of a third variable (class level proportion of students with SEN). Regarding this cross-level interaction, the results suggested that the within-class proportion of students with SEN and test scores slope was different for student with and without SEN.

In this study, it was possible to use longitudinal data with two data collection points. However, proper longitudinal models require data to be obtained for at least three occasions (Byrne, 2011); hence, individual growth or development over time was not studied here. Nevertheless, the prior performance was treated as a covariate in both models to control for the pre-existing differences in performance, and also to decrease the variance in later achievement; and thus, compute more precise estimates for class level variance (Blatchford & Mortimore, 1998; Konstantopoulos & Traynor, 2014; van Hek, Kraaykamp, & Pelzer, 2017). Furthermore, in general, due to a 'regression toward the mean' phenomenon, the scores from a second measurement will be closer to the mean, that is, students with high scores in the first measurement tend to move down on the second measurement while the students with low scores in the first measurement will move up (Allison, 1990). Thus, instead of calculating the gain scores, the first measurement was treated as an explanatory variable and the second measurement as a dependent variable in both models. It is clear, of course, that the test scores are not fully reliable measures of pre-existing conditions (Slavin, 1990) as we cannot assume that student sorting occurs only according students' prior academic achievement (e.g., Kupiainen, 2019; Paufler & Amrein-Beardsley, 2013). Nevertheless, when explaining the differences in student performance, the prior performance is typically an important factor, and according to Rivkin and his colleagues (2009), it often accounts for about half the variance. In Study I, the test performance in earlier grades accounted for slightly over 30% of the variation in later performance, and in Study II, a slightly under 40%. In both models, the prior test performance was also a relatively strong predictor of later performance, but

the effect slightly decreased after other variables were added to the models. According to Paufler & Amrein-Beardsley (2013), controlling for prior achievement helps to "level the playing field", particularly when the non-random student placement is studied.

3.5.2 Quasi-experiment using propensity score matching

In an ideal world of educational research, the effects of educational placement would be studied with a real experimental study—students (and teachers) would be randomly assigned to classrooms of equal sizes. Then we could assume that any class is as likely as any other class, and any teacher is as likely as any other teacher, to be assigned any student who is as likely as any other student to have similar background, school achievement, motivation, need for support and so forth. Briefly, the aim of random assignment is to make the probability of any observable differences among groups equal at the onset of any study (Gersten et al., 2005; Paufler & Amrein-Beardsley, 2013). When done correctly, randomization results in experimental and control groups that possess similar characteristics (Gersten, Jaynthi, Santoro, & Newman-Conchar, 2018). As the random assignment mainly eliminates pre-existing differences between groups and enables to control for extraneous influences that might bias the observations, in randomized experiments any causal inferences are expected to be more truthful (Shadish, Cook, & Campbell, 2002). Historically, experimental studies have been the only approach for estimating true treatment effects and making causal inferences (Lane, To, Shelley & Henson, 2012). In other fields such as psychology, medicine and public health, policy decisions are usually based on experimental designs with random assignment (Harker & Tymms, 2004). Despite its advantages, random assignment in the school world is rarely realizable or ethical. Instead, quasi-experiments can be created (Gersten et al., 2000).

Quasi-experiments lack random assignment (Shadish et al., 2002), and quasi-experimental studies can never completely replace true randomized experiments (Gersten et al., 2000). However, these alternative methods can also have desirable features—study conditions may be more representative of the real-life settings than in randomized experiments (Luellen, Shadish, & Clark, 2005). Furthermore, they can rule out the Hawthorne effect. That is, students participating in the experiment might act differently than they normally do because they know they are being studied (Krueger, 1999; Schanzenbach, 2010). One of the strengths of quasi-experimental approaches is that the participants are unaware that they are being studied, so Hawthorne effects are unlikely.

In true randomization, the participants (students), would have an equal probability of being assigned to either a special (treatment) or regular (comparison) class. As a result, classes could be compared with one another because any systematic initial differences would be controlled through the

experimental design whereas quasi-experiments are subject to participant selection: In this study, school-level decisions on student allocation introduces bias when the differences in the placement effect between groups are compared. Hence, groups may not be comparable at a baseline as non-randomized groups may systematically differ from one another and it can lead to a biased estimation of effect when these differences in the likelihood of group assignment have not been taken into account in the research design (Rosenbaum & Rubin, 1983).

The propensity score matching (PSM)[3] method accounts for this problem by using regression techniques to predict group assignment from theoretically relevant covariates and thus makes possible to match participants on these predicted scores (i.e. propensity score) (Rosenbaum & Rubin, 1983). In short, it is a mathematical approach to causal inference that uses participant's probability of group assignment to match participants between groups (Becker & Ichino, 2002; Lane et al., 2012). Thus, a propensity score is the conditional probability that a person will be in one condition rather than in another (e.g., get a treatment rather than be in the comparison group) given a set of observed covariates used to predict the person's condition (Rosenbaum & Rubin 1983). To oversimplify, the idea is that people who have the same propensity score but who choose (or are chosen) to be in different conditions are nevertheless comparable because the distributions of their covariates are in balance (Luellen et al., 2005; Stuart, 2010).

In Study III, a PSM was used to create statistically equivalent experimental (special class placement) and comparison (regular class placement) groups. For assessing the treatment effects using real-life data sets, Cuong (2013) has suggested that it is worth trying to match the non-participants with participants using all possible observed variables (see also, Thoemmes & Kim, 2011). In this study, Tier 3 students placed in regular classes were matched with Tier 3 students placed in special classes based on the closeness of the propensity score. Theoretically relevant pre-treatment variables are used to derive probabilities of group membership which are then used to match participants in treatment and comparison groups such that both groups have equal or likelihoods for the group membership (Lane et al., 2012). Potentially relevant covariates are those expected to affect treatment selection and outcomes (Luellen et al., 2005). The match rests on an assumption that propensity scores are free from hidden bias and that relevant covariates have been included in the model (Lane et al., 2012). Students in Study III were matched based on nine covariates measured at the beginning of seventh, soon after they were assigned to their lower secondary education classrooms. Following Gersten's and his colleagues' (2000), recommendations, student age, gender, SES, and achievement in cognitive tasks were considered. The idea is that, once matched on relevant covariates, any differences between these two groups

[3] There are different opinions as to whether PSM can be regarded as a quasi-experimental method. However, according to Holmes (2014), PSM is considered a quasi-experiment in this study.

should reflect the true treatment effects in the population and the differences can be interpreted as being similar to the effects of randomized designs (Lane et al., 2012). To put it in one question:

> "What is the expected effect on the outcome of Tier 3 students' cognitive outcomes and learning motivation if they were randomly assigned to special and regular classes?" (Adapted from Caliendo & Kopeing, 2008)

PSM is a fairly new, yet useful statistical innovation in educational research. In this study, I have exploited the pioneering work of Morgan, Frisco, Farkas, and Hibel (first published in 2010 and republished in 2017). In their research, the PSM has been utilized to assess the effectiveness of special education services comparing students receiving special education services to students not receiving such services, thus, the services have been considered as the treatment. Contrary to prior research design, in this study all students received support services at the same support level (Tier 3), thus the placement in a special class is treated as a treatment as it separates the students into two distinct groups. This is based on the hypothesis that the instruction received in special classes is distinctly different from the instruction provided in regular classes (Kauffman & Pullen, 1996). Furthermore, it can be argued that special and regular classes differ in terms of their size and composition, and in teacher qualifications (Zigmond & Kloo, 2017). However, in this study I have not taken a stand on whether placement in a special class is a preferable solution for Tier 3 students. However, to create statistical treatment and comparison groups, one group has to be the treatment group and the other the comparison group.

The data in Study III consisted of 860 Tier 3 students, 447 of them assigned to special classes and 413 to regular classes. Using PSM, 134 matched pairs were obtained using nine covariates to predict the placement. In practice, this means that for each student in the special class, a student in the special class that is as similar as possible in terms of their propensity score was identified. Like all probabilities, a propensity score ranges from 0 to 1. The placement in a special class (treatment) was coded as 1 and the placement in regular class (comparison) as 0. A propensity score above 0.50 meant the person was more likely to be selected for the treatment rather than the comparison group, and a score below 0.50 meant the opposite. The mean for propensity score was 0.47 for groups of students placed in special classes and 0.41 for groups of students placed in regular classes. Cases without a match were excluded from further analyses because they were considered to be outside the region of common support (Stuart, 2010; Thoemmes & Kim, 2011). According to Caliendo and Kopeing (2008), comparing the incomparable must be avoided, that is, only the subset of the comparison group that is comparable to the treatment group should be used in the analysis. Hence, the overlap and the region of common support between the treatment and control groups was checked (Stuart, 2010; Thoemmes & Kim, 2011). There were no

significant differences in any of the covariates between the matched and non-matched students and as expected, students placed in special classes showed a higher propensity of being placed in special classes than students placed in regular classes.

However, there was a substantial overlap of the distributions and as illustrated in Figure 6, the histograms before matching on the left do not differ much. Furthermore, the histograms after the matching on the right are similar. In sum, the matching was successful.

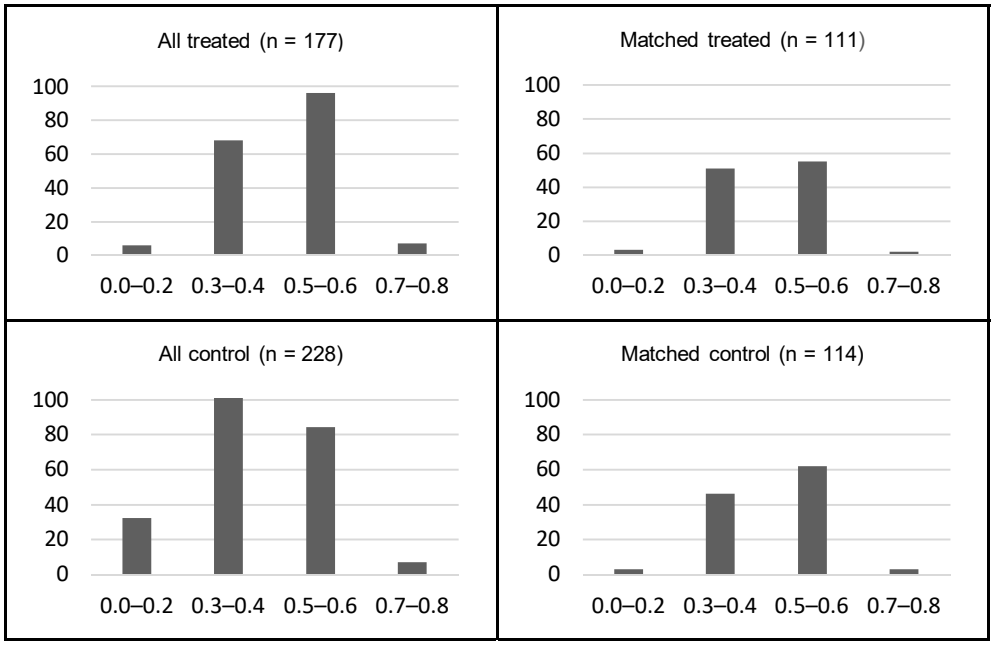

Figure 6 Histograms of propensity scores before and after the matching (X-axis: propensity, Y-axis: density)

Study III also operated with the effect sizes as the different experimental designs have to be accompanied with effects sizes (Gersten et al., 2005). Effect sizes were estimated with Cohen's *d* because the sample sizes were reasonable (Cohen, 1988, p. 477–478; Turner & Bernard, 2006). In general, effect sizes in the .40 or larger range are often considered to be the minimum levels for educational significance (Gersten & Smith-Johnson, 2001).

3.5.3 Research ethics

The present study has followed the ethical standards required of scientific research by following the ethical guidelines for the responsible conduct of research provided by the Finnish Advisory Board on Research Integrity (2012), and by following the ethical principles of research in the humanities and social and

behavioral sciences proposed by the National Advisory Board on Research Integrity (2019).

The primary education data (Study I). The study was conducted on an assignment from the Education Department of the city as an assessment study of the effectiveness of education, it was part of the education provider's duty to evaluate the education it provides and its impact (Basic Education Act 628/1998 21§). Thus, the permission for the data collection was obtained from the city administration. As the city administration had the authority to decide on the consent processes, the parents were informed of the assessment through the Education Department, securing the agreement of all the participating students. Further approval by an ethics committee was not required according to the local and national guidelines. The results of the original studies were reported to the municipality and to the participating schools.

The lower secondary education data (Studies II and III). The ethical approval for the data collection was obtained from the Ethical Committee of the National Institute of Health and Welfare. In all municipalities, the educational authorities adopted the study as a part of their internal evaluation of educational efficiency and its impact (Basic Education Act 628 628/1998 21§), and the permission to enter the schools was obtained from the city administration of each municipality for both data collections. Two municipalities required parents' permission for their children to participate in 2011 and one in 2014. In the other municipalities, parents were informed of the study as it was seen as an appropriate procedure by the city authority. The results of the original studies were reported to the municipalities.

In both studies, teachers conducted the data collections according to standardized written instructions and students completed the assessment tasks during ordinary school days. Participating students were assured that all information provided by them would be kept confidential and would be used for research purposes only. Both data sets were pseudonymized. The names of the participating students were initially included in the data in order to combine data from different data collection points. Subsequently, the names were replaced with artificial identifiers, that is, with unique numbers. Furthermore, the names of the municipalities, schools and classes were replaced with unique codes. The school codes were used to separate the different schools in the data, and the class codes maintained the nested structure of the data.

Both assessments (in primary education and in lower secondary education) were low-stakes test for the students and also for the schools. This means that they did not have any consequences for students and schools participating the data collection. The tests were not based on accountability policies, and no results were reported in a manner that how well schools were performing relative to each other (c.f., Smith & Douglas, 2014).

All research questions were answered with the most appropriate research methods, and all the analyses were reported in detail in the original articles. Furthermore, the limitations concerning the data and the analyses are listed in every study and in this overview. Thus, the results have been communicated in an open and responsible fashion in order to disseminate the results complying the responsible conduct of research.

3.6 Overview of the original studies

3.6.1 Study I

In Study I, the focus was on class size in regular classes in primary schools. In the study, the assumption that class size would positively predict primary school students' test scores due to the practice of placing students with SEN in smaller classes was tested. The hypothesis entailed that class size could be used as a means of support for Tier 2 and Tier 3 students. More specifically, whether class size has an impact on the student performance was analyzed after the initial differences between students were controlled for and whether the pattern would similar for students receiving Tier 2 or Tier 3 support. The data came from a longitudinal 'learning to learn' study conducted in a large municipality in Finland. The data used in the analyses consisted of 869 students (Tier 2 students n = 69, and Tier 3 students n = 36), from the fourth to the sixth grade. Learning to learn tasks were used to measure the performance. Items that were same in both data collections were used in the analyses. In this study, the cognitive items were referred to as to thinking skills. To test the hypotheses, a two-level regression model was specified to divide the variance in test scores into student and class levels. In addition to the regression coefficients, the bivariate correlations were studied.

In the first hypothesis, it was assumed that there would already be a positive correlation between class size and performance in the fourth grade and that Tier 2 and Tier 3 students would study in smaller classes. Descriptive statistics showed that on average, Tier 2 students were in classes of 21.49 students, Tier 3 students were in classes of 17.97 students, and other students were in classes with 21.91 students. Modeling the data confirmed the hypothesis partly: the positive correlation of class size and performance was statistically significant. However, only for the Tier 3 students had class size reduction been used as a means of support.

The second hypothesis tested the assumption that the gap between smaller and larger class increases over time, as a higher initial level of performance often predicts better later acquisition of the same skills and as students with SEN tend to fall behind in their development. According to the model, both Tier 2 and Tier 3 support was related to lower initial performance and they also predicted negatively the test performance in sixth grade. At the individual level, student's

SEN status was related to the lower initial performance and the gap increased during the follow-up. However, at the class level, the proportion of students with SEN in the class predicted later performance positively.

In the last hypothesis, it was assumed that regardless of the general tendency of class size being positively related to performance and its development, for students with SEN, the prediction would go in the opposite direction and studying in smaller classes would be helpful to them. There was no interaction between class size and Tier 3 students, and a very weak interaction with class size for Tier 2 students. However, due to the relatively small sample size, the findings could not be confirmed.

In all, it could be concluded that class size was only related to initial differences and not to the development of performance. It could also be concluded that the assignment of Tier 3 students to smaller classes could be one way to manage increasing student heterogeneity in regular classrooms.

3.6.2 Study II

The aim in Study II was to investigate the class composition effect is in terms of the proportion of students with SEN in regular class. I examined how the proportion of students with SEN in regular classes was related to the student-level and class-level performance. Also studied was if the effect was different on students with SEN than for students without SEN. The data (N = 5368) were drawn from a longitudinal assessment study conducted on students at the beginning and at the end of their lower secondary education in a Finnish metropolitan area in 14 municipalities. Data consisted of 256 Tier 2 students and 144 Tier 3 students and 4 968 students without SEN status. Learning to learn tasks were used as a measure of performance and referred to as cross-curricular competencies. A two-level regression model was specified. Furthermore, cross-level interactions were also defined in order to study the compositional effects at the student level more closely.

First, it was concluded that in classes with students with SEN, the average performance level was lower than in other classes. At the beginning of the modeling, class-level SEN status was treated as a dichotomous variable, i.e., whether there was one or more students with SEN in class or no students with SEN in class at all, and it did not predict the ninth grade test performance. This indicates that the presence of occasional Tier 2 or Tier 3 students in a class does not lower the overall performance level of a class. However, when the proportion of students with SEN was treated as a continuous variable, it predicted negatively, though weakly, the test scores in the ninth grade. Moreover, the results indicated that the higher the proportion of students with SEN, the lower the ninth-grade class-level test scores. Modeling also revealed that when the proportion of students with SEN in class reached 50% of the student population, the students

without SEN performed at the same level as the students with SEN. However, as the proportion of students with SEN was not normally distributed, rather it was strongly negatively skewed, it was divided into quartiles. Even then the trend was visible. When the proportion of students with SEN in class exceed 30%, the performance level of the students without SEN declined. The same was concluded adding a random slope to the model for the class-level proportion of student with SEN on students' SEN status. It revealed that the more students there were with SEN in class, the stronger was the effect.

Finally, by adding a cross-level interaction to the model, it was possible to study how the class-level proportion of students with SEN affected the student-level performance. It was concluded that the proportion of students with SEN in class had a slightly different effect on performance for students with SEN than for students without SEN. Students with SEN performed at the same level regardless of the proportion of other students with SEN in class. However, the performance of students without SEN in classes with students with SEN was slightly lower than their peers in classes without students with SEN in the ninth-grade assessment, even when the initial performance differences were taken into account.

The results yielded that purposeful sorting of students into classrooms is one way to manage student diversity in schools, especially concerning students with SEN. The results may be explained more by the classroom assignment practices than by actual compositional effects—there seems to be a tendency to create more homogeneous classrooms as less-achieving students without SEN are placed with students with SEN. However, this can increase the between-class differences in Finnish schools.

3.6.3 Study III

The aim in Study III was to investigate the differences between Tier 3 students assigned to special or regular classes in the lower secondary education level. Specifically, the differences were studied in curriculum-based Finnish and mathematics tests, in grades, and in self-reported goal-orientation scales. Data were drawn from a longitudinal assessment study conducted on students at the beginning and at the end of lower secondary education in a Finnish metropolitan area. The data consisted of 860 Tier 3 students, of which 447 were assigned to regular classes and 413 to special classes. Propensity score matching (PSM) was used to match the students according to their propensity to be placed in special classes, and 134 pairs were produced. PSM is a quasi-experimental method that creates a statistically equivalent experiment and control groups and thus, enables the effects of placement to be detected. The matching was done using students' seventh grade test achievement and various background variables. The differences between the two groups were studied with t-tests and one-way and two-way ANOVA.

There were no major initial differences between students with SEN placed in special (n = 134) or regular classes (n = 134) at the beginning of the seventh grade. Furthermore, students assigned to special classes did not differ in curricular Finnish and mathematics tasks in the ninth grade when compared to their closely-matched peers in regular classes. However, students in special classes had a higher GPA than students in regular classes. The separate grades were also studied: students in special classes had average higher grades in mathematics and Finnish. The possible differences in assessment practices due to more individualized curricula for some students did not explain these differences.

In addition to cognitive outcomes, the learning motivation was also examined. Students in special classes scored higher in mastery-extrinsic orientation than their matched peers in regular classes Furthermore, they had a higher performance-approach orientation. In general, the effect sizes were small, yet, grade differences yielded medium effect sizes.

To sum up, students in special classes did not perform poorer in curricular tasks at either the beginning or the end of their lower secondary education. However, they earned higher grades in these subjects. Furthermore, given the higher scores in self-reported mastery-extrinsic orientation and higher performance-approach orientation, the findings may indicate that in smaller special classes, success in schoolwork is emphasized more than improving the level of competence. In some respects, these results could be the result of higher academic self-concepts in more homogenous classes because it is easier to outperform students at the same ability level.

It is evident that despite the general tendency to place all students in regular classes, there is a continuous group of students that will still be taught in special classes and that the placement can affect learning motivation and even school achievement measured by grades.

4 General discussion

In this chapter, I will reflect on the results flowing from this study in relation to the research questions set for this dissertation. I will start by outlining the main findings in the original studies and discuss them in relation to the literature in the field. After this, I will consider some of the gaps in this study. That has been done in two parts. As the data and research methodology set always limits for what can be concluded based on the findings, methodological solutions are reflected on briefly. Firstly, as every piece of research is a balance between compromises, I will discuss the overall limitations of this study that should be kept in mind when interpreting the results. Secondly, I will focus on the methodological aspects and ponder on what can be concluded based on statistical inferences. It is said that useful educational research should focus on factors that significantly influence the quality of classroom teaching and learning. Furthermore, such research should be responsive to the concerns of teachers and those whom they serve, and should be helpful to policy-makers in their practical decision-making (Pedder, 2006). Thus, some implications and conclusions are provided, as the basis for a wider discussion. In addition, I will look at the future and make some suggestion about what should be done next.

4.1 Main findings of the studies

The aim of this study was to detect the class placement mechanisms in terms of students with SEN in the Finnish context. This was done by examining the class size and composition effects on students with SEN and without SEN (Studies I and II), and by testing the placement effect on students with SEN (Study III).

According to Dreeben & Barr (1988), schools can be seen as providers of curricular instruction, and by subdividing populations of students they can make instruction appropriate to the capacities of students—in their view, the main task of school is to provide appropriate instruction to a large and diverse clientele in classes of workable size and composition. This study treated students in clusters (classes in schools) with a hypothesis that students can expected to have different levels of performance depending on the class they are assigned to. It seems that most issues important to both general and special education are, at the same time, contemporary and perpetual. Naturally, the terms can vary and they can be reframed in different periods of time, yet the main questions stay unchanged (Kauffman, Nelson et al., 2017, cf., Eliot, 1933; Postel, 1937). Questions like what are the school- and class-level factors that can enhance student learning in a way that everyone can reach their full potential? Or, what is the most desirable placement for students with SEN?

Admittedly, this study cannot solve the whole puzzle either. However, the aim with this study was to provide much needed evidence in the Finnish context and thus, offer a direction for future research. The perpetual issues specifically in special education are: who should receive special education, how should students be identified, and where should they be taught (where should they learn)? (Kauffman, Nelson et al., 2017). This study focused on the second question. It is clear that policymakers, teachers and parents are interested in identifying the factors that can enhance academic development at school for all students. As the placement of students with SEN concerns all students, not just those with SEN, the question is more than special educational; it covers the whole educational system, as Studies I and II have indicated.

One of the main objectives of this research was to serve as an initial study on the class placement of students with SEN and on the class compositional questions in Finland. If the findings were to be summarized in one sentence, it could be said that class placement matters for all students, not only for students with SEN. Faced with a diverse student population, schools divide students into relatively homogeneous categories so they can deal with different groups of students in different ways (Gamoran et al., 1995). It is evident that schools must have some means of differentiating between students to take into account the variation between them. Class composition is a powerful solution (Thuneberg et al., 2015). Based on previous research, in Finland, it tends not to matter that much which school a student attends (OECD, 2016b). Yet, in light of this and few other Finnish studies, it matters a lot which classrooms in that school a student is in (Kupiainen, 2019, Thuneberg et al., 2015). Furthermore, it is not the class size that makes the difference, rather it is the composition of the class. Consequently, in addition to the number of students in class, other aspects in class composition should always be considered when studying the class-level effects.

To sum up, student placement is far from neutral. According to US-based study on the student placement practices with principal survey data, "student placement can make or break a student's learning" (Paufler & Amrein-Beardsley, 2013). Furthermore, quoting another principal: "Anything done randomly will get random results. If assignment of students is done strategically with a goal in mind (student success) then there is a higher likelihood of meeting that goal". As concluded, students with SEN face a wider continuum of placement options and require case by case consideration (Jahnukainen, 2015). In the next section, the main findings of this study are discussed, both in the light of the previous research, and the common understanding of the Finnish education.

4.1.1 Class size as means of support

In Study I, the focus was on class size effects in regular classes in primary schools. According to the literature, small class sizes may matter more in the early grades, when students start to socialize at school and form work habits; in the other words, they learn how to be at school (Ehrenberg et al., 2001). Based on the vast body of research, it is known that class size effects are not easy to discern and that the findings have been highly ambiguous, also based on a previous Finnish study by Kupiainen and Hienonen (2016). Therefore, the focus was especially on students with SEN. In line with international research, it was hypothesized that the lower-performing and disadvantaged students could benefit from smaller classes more than others (e.g., Blatchford, Basset et al., 2011; Blatchford, Goldstein et al., 2002; Blatchford & Mortimore, 1994; Finn & Achilles, 1990; Finn & Achilles, 1999; Hargreaves at al., 1998; Molnar et al., 1999). The effect of class size was not statistically significant when all the students were considered. The effect was also studied by focusing only on students with SEN. Rather unexpectedly, the class size effect on Tier 3 students was negligible whereas there was weak effect on Tier 2 students. Thus, the study gave faint indications that students at Tier 2 level would benefit from smaller classes. However, the data lacked the power to study this effect further.

The study partly confirmed the phenomenon in Finnish schools that students with SEN are placed in smaller classes, and that class size is used as means of support for these students. Based on the data, Tier 3 students were placed in smaller classes, whereas the average class size for Tier 2 students did not differ from the average class size of students without SEN. However, these students were the ones who might have benefited from smaller classes.

It turned out that no automatic effects on student performance occur in smaller classes. There are multiple explanations that could explain the findings of this study. It seems plausible that the effects of class size are mediated by an array of instructional and peer processes (Blatchford & Martin, 1998). The number of students in a class may enhance or impede learning, yet better learning results is the sum of many processes. Many factors influence student learning and we cannot expect that reducing class size will automatically and directly affect student learning. Rather, according to Wilkinson and his colleagues (2002), reducing class size merely increases the probability that the environment can be structured to increase learning, motivation and a positive classroom climate.

In education policy, there has been great belief in class size and its effects, partly because it is one of the simplest variables for policymakers to manipulate (Ehrenberg et al., 2001). Maybe we also should be responsive to the possible benefits for student learning that arise in large or small classes and therefore it is important to guarantee resources for schools to adopt more flexible approaches to allocate students to classes of different size for different teaching and learning purposes and needs.

4.1.2 Students with SEN in regular classes

Moving from the size of a class to the composition of a class, as students with SEN are increasingly placed in regular classes, some concerns have been raised. First, there has been discussion about how this affects the overall performance level of a class. In Study I, there was a slight positive effect: the proportion of students with SEN in class predicted the class-level performance positively. This suggested that support could have been adequately provided in those classes, and it also lent some support to findings that additional support provided in regular classes for students with SEN could be of benefit to the whole class (Cole, Waldron, & Majd, 2004; Ruijs, 2017; Thuneberg et al., 2013). However, Study II yielded opposite results. The average performance level in regular classes with students with SEN was lower than in regular classes without students with SEN at the beginning and at the end of the years in lower secondary education. These findings must be interpreted with caution. The results do not suggest that teaching or learning would be inferior in classes where there are students with SEN or that students with SEN would impede the learning of other students. One must keep in mind that the classroom processes were not under scrutiny in this study. However, some speculations can be presented in light of this previous research. Firstly, the lack of adequate resources is under constant debate, especially when students with SEN are placed in regular classes (e.g., YLE 18.2.2019). The data in Study I were collected in primary education whereas the data in Study II were obtained from lower secondary education. This can partly explain the different results. Classroom teachers in primary education have a different educational background from subject teachers in lower secondary education. This can also affect the preparedness to provide support in classroom.

The findings in Study II can partly be explained by the lack of adequate support in classrooms to allow for the heterogeneity of all learners. Secondly, because of this hidden tracking, teachers might simply teach the lower-performing classes differently by lowering the overall attainment level of the class independent of the presence or absence of students with SEN. This is also in line with the research on teacher expectancy. According to Goldenberg (1992), a teacher who has low expectations is less likely to present advanced material and more likely to provide less demanding material. Another explanation might be that differential teacher practices are employed depending on the composition of the classes (Wilkinson et al., 2000). Pedder (2006), argues that teachers treat students differently according to the expectations they have for their potential learning. How a teacher is prepared to work and what resources are available in classrooms are admittedly linked to teaching practices (Szumski et al., 2017).

The average proportion of students with SEN in regular classes was 13%, however, it ranged from 0% to 50%. The modeling revealed that the more students with SEN there were in class, the stronger the slight negative effect of the presence of students with SEN in class was. In a way, results of this kind may be expected.

There more students with various needs in one classroom there are, the more challenging it becomes to meet the need of every student. In light of this, it becomes extremely important that the support resources follow the student to the classroom.

It has been commented that even though students without SEN constitute the majority in regular classes, most studies usually focus on students with SEN (Fletcher, 2010; Lindsay, 2007). In Study II, the effects on students without SEN were also examined. The findings revealed that students without SEN, placed in classes in which there were students with SEN, performed less well than their seventh grade peers in classes without students with SEN. It is evident that schools assign students to classes with the intention of managing academic diversity among students by reducing the heterogeneity within instructional groups. Dividing students into more homogenous sub-groups appears to be a logical and sensible means of responding to student variability and organizing a student body with diverse skills and needs. It could also be that this allows teachers to tailor their instruction to students' abilities and support needs. It is easy to understand the main logic behind assigning students this way. However, there might be some unintended consequences. Gamoran and his colleagues (1995) noted that if students are divided on the basis of academic criteria, they also tend to be divided by socioeconomic characteristics. Moreover, if the idea is to attempt to provide appropriate instruction for each group of students, it can result in a situation in which students in lower-performing classes tend to receive inferior instruction compared to higher preforming classes. This kind of student allocation can imply hidden or informal ability grouping which can for one explain the considerable differences between classes in Finnish schools.

One of the findings that needs to be discussed in more detail is that the proportion of students with SEN had a positive effect on students with SEN. Previous research by Ruijs, Peetsma & van der Veen (2010) found no such differences. The results of this study may indicate that when a couple of students with SEN are taught in the same regular class, they may benefit from it. This may also imply that on these occasions, the support had been provided successfully. However, the data did not enable an explanation of the findings. It is always possible that students with SEN may have higher achievements, because they can learn from more able students in regular classes (Ruijs & Peetsma, 2009). At the same time, when students with SEN seemed to benefit from these practices of class allocation, students without SEN appeared to lose ground. A study by Huber et al. (2001) yielded similar findings. Furthermore, studies on spillover effect by Fletcher (2009; 2010) found evidence that having a classmate with an emotional problem is associated with lower test scores in reading and mathematics. This leads to concluding and repeating what was said at the beginning of this chapter: classroom allocation and class composition matters for all learners.

4.1.3 Regular or special class?

The focus in Study III was solely on Tier 3 students and on their classroom placement. This study also added to the other two studies by considering learning motivation in addition to performance.

There were no major initial differences between Tier 3 students placed in special or regular classes when different background variables at the beginning of the seventh grade were studied. However, schools and municipalities have different policies for assigning students to classes and they cannot be detected in comparisons based on means.

Findings by Dessemontet et al. (2012), Peetsma et al. (2001) and Rea et al. (2002) have favored placement in regular classes when various academic outcome measures have been studied. However, the data in Study III did not indicate any differences in cognitive tasks measured by test scores in curricular and cross-curricular tasks among Tier 3 students placed in regular or in special classes. Students in regular classes did not perform poorer nor did they outperform their counterparts in special classes. This may indicate that student placement practices have been suitable as no performance differences as such were detected.

Yet, the picture was slightly different when the school achievement measured with school grades at the end of the ninth grade were studied. Students in special classes had higher overall GPA and higher grades in Finnish and mathematics. Prior research had indicated that students with learning disabilities earned significantly higher grades in regular classroom (Rea et al., 2002). It is clear that grades are problematic measures for comparing performance differences. However, the grades have some relevance to the students as students are informed about their performance by their grades. Furthermore, grades are important in determining educational progress as students are compared in terms of the grades at the end of the ninth grade when they apply for entry to upper secondary education. According to the National Core Curriculum, students are not to be compared when grades are given (FNBE 2016). However, there are indications that instead of national criteria set for different grades in different subjects, teachers adjust their grades to the overall competence level of the class (Ouakrim-Soivio, 2013). In line with the findings of Ouakrim-Soivio, it can be argued that corresponding performance in special classes may produce higher grades in special classes than in regular classes. Previous research has also indicated that teachers are more prone to allowing the lower performance to be compensated with extra effort when assessing students with SEN (Rojewski, Pollard, & Meers, 1990). As the students did not differ in Finnish or mathematic tasks but received higher grades, it may lead to unrealistic grades which may be troubling at the upper secondary education.

Regarding the learning motivation, students in special classes had higher mastery-extrinsic and performance-approach orientation. In other words, students in special classes had aspirations to have higher grades, to succeed better than

their peers, and to show their abilities to others. At the same time, there were no differences in mastery-intrinsic orientation between the two settings. In general, learning motivation has to be considered because it has been shown that it plays a part in classroom learning for students with SEN (Botsas & Padeliadu, 2003; Schwab & Hessel, 2015). With this in mind, the findings may indicate that in smaller special classes, success in schoolwork is emphasized more than gaining competence. Encouraging self-improvement in general can be positive for all students while encouraging comparison among peers may be less positive for lower achieving students in regular classes (Patrick, Kaplan, & Ryan, 2011). On the other hand, students in special classes compare themselves to students with a similar performance level, which can lead to a more positive self-perception concerning school tasks (Bakker et al., 2007; Belfi et al., 2012); and thus, higher academic self-concept in special classes (Törmänen & Roebers, 2017). This can lead to concluding that encouraging comparison and striving to succeed more than others could function better in more homogenous special classes. There is some evidence that the goal orientation would function as a moderator for the Big-Fish-Little-Pond-effect (BFLPE); that is, students with high endorsement of extrinsic goal orientation would experience stronger BFLPE (Cheng et al., 2014; Wouters, Colpin, Van Damme, & Verschueren, 2013).

To conclude, in some respects, the higher intrinsic-mastery and performance-approach orientations could be the result of higher academic self-concepts in more homogenous classes because it is easier to outperform students at the same ability level. In addition, if a student then manages to succeed in school and outperform classmates, it can raise their academic self-concept. The higher grades could partly be explained by the higher performance goal structure observed in special classes, which in turn may lead to seeking recognition and extrinsic rewarding. Grading may be used as an incentive to induce students to engage with certain learning goals, and these goals can be rewarded with higher grades (Ames, 1992). However, this is mostly a post hoc argument, as the reciprocal effects between self-concept, performance and goal orientation was not in the scope of the present study. This calls for future research.

4.2 Limitations of the study

One advantage of this study is that it used the data of two well-designed, longitudinal large-scale assessment studies conducted in Finland. Two carefully collected data sets also made it possible to employ sophisticated multilevel models and quasi-experiment in this study. Nevertheless, there are limitations that should be considered when assessing the overall results of this study. Next, the major issues concerning the limitations are presented and then some minor points are provided.

Firstly, the use of secondary data sets some restrictions for this study. In an ideal world, data collection should come, if possible, after the models of interest have been planned (Lei & Wu, 2007). However, the investigation of class size, composition or placement has not been the primary purpose of either of the larger research projects. Analyses in this study have been carried out after data had been collected to fulfil other purposes. There are many variables that could have been included in the original data collection if they had been carried after writing the first draft of the research plan. However, it is unlikely that a doctoral researcher could collect such extensive data for the purpose of the doctoral dissertation. Considering the nature of the research questions and the statistical methods required to answer them, the best possible solution was sought. Moreover, as I had been involved in all data collections and many other stages of the original research projects, I was also familiar with data and their limits. Acknowledging the restrictions, all possible variables were included in the analyses.

Secondly, one feature of special education research that makes it complex is the variability and diversity of the participants. That is, the heterogeneity of participant characteristics poses a significant challenge to the research designs (Odom et al., 2005). Students with SEN may have very different special educational needs; naturally, different needs require different classroom placements and these placements may have different effects on different students (Fletcher, 2010). Unfortunately, the data do not inform about the types of diagnoses or ground for receiving support at different tier levels: this means that it was not possible to make any distinction between the different types of SEN (for example, a student with SEN in the area of learning difficulties or in the area of behavioral and emotional difficulties). To avoid a heavy workload for the teachers, this background information was not requested. Furthermore, collecting such individually based data would have been against the Finnish policy of avoiding diagnosis-based categorizing of students with SEN. However, I acknowledge that the type of SEN may influence placement decisions. Previous research has suggested that certain groups of students (i.e. high-incidence disabilities) are more likely to be placed in regular classrooms whereas students with significant disabilities are more likely to be placed in separate settings (Morningstar et al., 2017). Then again, a study by Ruijs (2017) indicated that distinguishing between different types of SEN did not change the effect of the presence of students with SEN in class. Furthermore, students with the most severe disabilities were excluded from the data collection as the assessment situation and the tasks would have been too demanding according to their teachers' views. In Finland, it would mean all the students studying according to functional areas (4.9% of all Tier 3 students, OSF, 2019). Naturally, if enough background variables would have been available for this research, a closer look at the representativeness of the Tier 3 students in the analyses would have strengthened this study.

Thirdly, the data did not contain information on how the support for students in classes was implemented in each school and class (cf., Gersten et al., 2005). There might be substantial differences in the arrangements across schools, which in turn could have different effects on students. Furthermore, the determinants of placement decisions may vary between schools and municipalities. However, all schools function under the same legislation, the National Core Curriculum, and to some extent, nearly the same set of budgetary constraints, though the local authorities can determine the use of the funds allocated by government (Pulkkinen & Jahnukainen, 2016). Nevertheless, there is variation across municipalities in how and where support is provided (Lintuvuori, 2019). It is important in future work to examine potential heterogeneity in effects by municipal-level policies that might shape the student assignment. Furthermore, when it comes to student performance, the differences between Finnish schools have traditionally been quite small (e.g., OECD, 2016b). Thus, school-level analysis was excluded, as there was no substantial variation to explain. However, in the future, school-level analysis could be included in the model and treated as a fixed effect.

In addition, data do not provide information on the amount of time the students with SEN are included in regular classes—in future, information on whether students were withdrawn to special classes for certain lessons, and for how long, is required.

In many observational studies, class size is not measured accurately because data about the actual class size in each classroom are not available. Instead, class size is frequently calculated as an administrative enrolment number by the teachers (e.g., student- teacher ratio) in each school (Konstantopoulos & Traynor, 2014). In this study, because the number of students in class was extracted from the original student lists, the real class size could be obtained. However, the data did not include the information on the time students spent in their class. In addition, especially in lower secondary education, class composition tends to vary slightly by subject area, different classes are staffed by a different teacher, and therefore the class composition can vary for each student during a school day. In short, class size is easier to define in primary education but more difficult in lower secondary education (Hoxby, 2000). Taken together, as a fifth limitation, the classroom nesting was more difficult to conceptualize in Studies II and III.

While classroom differences may be partly due to the sorting of students into different classes, they may also reflect differences in quality of instruction (Ehrenberg et al., 2001). More broadly, compositional effects can be explained by teacher effect as well as peer effects. However, the data did not include data on classroom practices or peer relations. Thus, the sixth limitation of this study is that only the effects of contextual variables were analyzed, but there was no examination of the processes through which classroom composition and placement affected the student performance. How different instructional practices in classrooms could impact or mediate the class size and class composition effects

was not examined. However, such observational data were well beyond the scope of this study, as well as beyond the data collection capacities of the researchers.

Some attention should also be paid to the LTL measures used in this study. The students in lower secondary education (Studies II & III) completed the same tasks at both stages of the data collections, and the possible retesting effect was not tested. However, since there was a two-year gap between the data collections, the retesting effect is quite unlikely. On the other hand, all items from grades 4 and 6 of the primary education data could have been included by applying item response theory (IRT) and linking them with IRT-scores (de Ayala, 2009). Furthermore, the possibility of the regression towards the mean should be mentioned (e.g., Bland & Altman, 1994). It is always possible that at the second measurement point, the highest-achieving students will obtain lower scores and the lowest-achieving will obtain higher scores, in this case, favoring Tier 3 students in the classes. Additionally, one option would have been the use of gain scores. However, the use of gain scores includes its own limitations, and for this reason they were not used in this study (e.g., May & Hittner, 2010).

Good research design always balances between compromises (Gersten et al., 2000). At the same time, one of the strengths of this research is the statistical methods. However, they also set some limits. Any statistical analysis, no matter how sophisticated, rests always on quantification. Even at their best, statistical models are simplifications and approximations of real-world phenomena (Lee, 200; McDonald & Ho, 2002). Every researcher must acknowledge the fact that there may be some essential processes and their effects which cannot be captured in this way.

In Studies I and II, hierarchical models were employed to respect the nature of the data and the phenomenon. However, in Study III, more conventional single-level methods were used as the sample size was not sufficient for more rigorous multilevel models (Gersten et al., 2000; Lee, 2000; Maas & Hox, 2005). Furthermore, the amount of variance explained by the class level (ICC) was too trivial for multilevel methods to be considered. The biases described in Chapter 3.5.1, occurring when multilevel data is treated with single-level methods, have to be mentioned. As the analyses had to be done at the student level, a class measures were appended onto each student in a particular class. Admittedly, this assumes that the performance and learning motivation of all students in the class were influenced identically by the classroom assignment. The methodological choices were made acknowledging these restrictions, and at the same time, planning a new research design in which the data collection was targeted primarily at the research questions.

The last restriction of this study is the limited geographical area covered by the data collection which focuses only on the southern parts of Finland. Both sets of data used in this study represent the Helsinki Metropolitan area, and even though it is the largest urban area in Finland, it does not represent the whole country. This

has to be mentioned when considering the generalization of the results. On a broader scope, researchers in special education and education in general, face always particular problems in comparison to the natural sciences and they must deal with local conditions that limit generalizations (Berliner, 2002).

The use of rigorous research methods will enhance the quality of research in special education, and thus, will hopefully improve our understanding of what works for whom and in what context, and in the end, improve the education we provide to students (Gersten et al., 2018). Nevertheless, it is important to remember that no research design is perfect. There are always trade-offs and considerations related to the alignment of research questions with appropriate data and methodologies (Gersten et al., 2018). However, one has to start somewhere. I hope this will be reserved as a preliminary study in the Finnish context, and in addition to the results, I hope this points out some critical aspects in this kind of research and helps to design new, enhanced data collections.

4.3 Methodological reflections

One of the main strengths of the present study has been the secondary data utilized throughout the study but at the same time, this is also a challenge. In most cases, it was possible to choose the most appropriate methodological solutions for testing the hypotheses posed. However, the use of secondary data also set some limits for the analyses. Furthermore, the complexity of the phenomenon under scrutiny increased the challenges for the analyses. The methodological solutions were described in Chapter 3.5 and the limitations for the study concerning the methodological aspects and data in Chapter 4.2. Here, some concluding marks and afterthoughts are presented, bearing in mind what should take into account in future research.

Respecting the nature of the phenomenon studied, two main methodological strategies were chosen: multilevel modeling and propensity score matching (Odom et al., 2005). However, it is clear that every methodology has its limits (Greenland, 2017). Furthermore, as well as being class placement and compositional effects, class size effects are hard to pin down. Multilevel models take into account the hierarchical structure of the data and thus, in Studies I and II, they were the main methodological approaches. However, when constructing models which involve compositional effects, model specification and predictor reliability should be taken into careful consideration. As Harker and Tymms (2004) put it: “It is only after detailed and careful work in an area by many researchers that we can be fairly sure that their models are suitably structured”.

As discussed briefly in Chapter 2.2.3, in terms of class compositional effect, it is however possible that the statistical procedures, which indicate the existence of a compositional effect, are misleading and are due to the statistical artefact referred to as the phantom effect (Harker & Tymms, 2004; Televantou et al.,

2015). Such an effect can appear in multilevel models as a result of unreliable data or poor model completeness. However, data used in this study were obtained from two carefully designed large-scale assessments with instruments and measures developed over the years. Furthermore, the use of longitudinal data in both models enabled the effects of class-aggregated variables after adjusting for the initial student-level differences to be estimated. In addition, some background variables were controlled for, to consider possible confounders, such as SES and age, before the final models were estimated. Moreover, overall sample sizes and cluster sizes were adequate (Gersten et al., 2000; Maas & Hox, 2005). Classes with fewer than 10 students present at the time of the assessment were excluded from the analyses, as their class-level results were not seen as representative.

Even though it was possible to utilize carefully considered data, some aspects are still worth discussing. When we look at the variables, the concept of class size is somewhat challenging even though it may seem to be an obvious and easily available measure (Blatchford, Basset et al., 2011). In this study, the class size was based on the student lists provided by the education providers and it is more accurate than the statistically calculated class size (see, OECD, 2019a). The same applies to the measure of class composition. According to Konstantopoulos and Traynor (2014), when the classroom which each student belongs to is correctly represented in the data, it makes it possible to overcome some of the shortages typical in class-level studies. Naturally, class compositional measures have a temporal dimension that is seldom available. The students in the class at any time may be different from the calculated class compositional measure. To obtain such detailed data, a carefully planned data collection that is preceded by the research hypotheses is warranted.

In general, multilevel models allow for to variables at the different level of hierarchical data to be taken into account. However, as Lüdtke et al. (2008) noted, the focus of student level measures is on a student level construct, and individuals within the same class are likely to have different true scores. This means that scores in LTL tasks for different individuals within the same class are not interchangeable, and that the individual level and aggregated variables do not always reflect the same construct. In the models, the class level variables for performance have been presented as means which may hide a part of the variation in a class. In the preliminary analyses, the standard deviation of each class was calculated from the individual scores as an indicator of class heterogeneity and entered into the model. The effect was non-existent and thus, it was excluded from the final model. Furthermore, in multilevel models, a variable can serve as both an independent source variable (an exogenous variable) and a dependent result variable (an endogenous variable) in a chain of causal hypotheses (Lei & Wu, 2007; Tarka, 2017). When schools use purposeful sorting of students into classrooms and class size is used as a means to manage student heterogeneity, it is worth considering whether the causality could go the opposite direction. Thus,

the variation in class size could be explained by the performance differences (see Hoxby, 2000). However, some of these concerns can be ruled out by the use of longitudinal data in this study.

There is always also a need for some caution with the model interpretation. Multilevel models are often correlational and thus, causal conclusions must be defined carefully (Hox & Bechger, 1998; Televantou et al., 2015). Relations in the models are hypotheses about directional influences (e.g., how the number of students in a class affects class-level performance) or causal relationships between multiple variables and they are not deterministic. This means that in this study, it should be noted that the detected relationships are correlational, and no causal inferences can be made. Rather, they only increase the probability that an effect will occur (Lei & Wu, 2007; Shadish et al., 2002). To put it more simply, statistical models do not magically transform correlational data into causal conclusions (Hox & Bechger, 1998). In addition, especially in educational sciences, many factors are often required for some effect to occur, and we rarely know all of them and how they relate to each other. It is nearly impossible to adequately control for all the factors affecting the studied phenomenon. Finally, in terms of the model interpretations, they can indeed improve causal descriptions, but they do not always explain causal relationships (Shadish et al., 2002). That is why qualitative data are also needed.

In the context of the present study, the above mentioned means that the class compositional features such as the number of students in class or the proportion of students with SEN in class can affect the performance level of the class but it does not mean that it automatically and necessarily always affects it in the same way. It only means that the more students with SEN there are in class, the greater the probability that the student performance will be lower than in a class in which there are no students with SEN. In addition, it only indicates that there is a relationship between these two variables but does not explain that relationship. It can partly be caused by the non-random student allocation and partly due the lack of adequate support sources for all students in class. Different classroom practices can mediate the effects and to explain these mechanisms, more research called for the aforementioned qualitative data.

In Study III, a quasi-experiment was created. It is clear that quasi-experiments can never substitute for a true random assignment experiment (Gersten et al., 2000; Shadish et al., 2002). Nevertheless, in this study, instead of backing down, in order to provide foundation for future research, the quasi-experiment was employed using propensity score matching. It is a relatively unknown, yet useful method. Much of experimental and non-experimental studies are influenced by the methodological traditions from psychology and medicine (Gersten et al., 2017), however, this study adopted a more policy-oriented research tradition originating from seminal work by Campbell and Stanley (1963).

Propensity score matching operates with the treatment effect by diving the data into treatment and control groups. One could always argue against the idea that the placement in a special class is seen as a treatment, especially when the data do not include any measures on what happens in the classrooms. However, as the special classes differ from regular classes in terms of size, composition and teacher qualification (Zigmond & Kloo, 2017), it is safe to say that the context in a special class is different from that in a regular class, and thus, can be handled as a treatment variable.

Furthermore, although with careful research design it is possible to establish equivalence on observed characteristics, one can never be completely sure if the groups are equivalent on other unmeasured characteristics. Naturally, one can never detect every possible variable on which the experimental and comparison groups may differ; it is always possible that an unknown variable(s) could be partly responsible for the results (Gersten et al., 2000). Thus, there is always a question if one group performed better on an assessment because the treatment or because the groups were different in terms of some unobserved characteristic. While selection bias cannot be completely ruled out in the absence of a true randomized experiment, with a set of background variables, the differences between the treatment and comparison groups were within recommended limits (Gersten et al., 2000). Moreover, quasi-experimental studies are considered worthy contributors, but only if they equate groups and control for pretest differences (Gersten et al., 2018). In this study, there were almost no pre-existing differences between the two groups, which strengthens the reliability of the results. Yet, there would have been options other than the propensity score matching method. In multi-level modeling, school-level fixed effects would have been one way to control for the non-random student allocation as well as the use of student weights (Hong, 2015). In future research, these different methodological options should definitely be compared.

To sum up, schools are incredibly complex organizations, impacted by forces from society, parents, policy makers, teachers, individual student characteristics. (Harker & Tymms, 2004). Under these circumstances it is surely quite impossible for any statistical technique to be able to explain it all. Another thing is that the perfect statistical model does not exist. However, quoting Shadish et al. (2002. p. 457): “The purpose of experiments is not to completely explain some phenomenon; it is to identify whether a particular variable or small set of variables makes a marginal difference in some outcome over and above all the other forces affecting that outcome”. It is clear that there is a great need for multiple approaches, involving qualitative and quantitative methods that explore the context of classroom deeper.

4.4 Conclusions, implications and future directions

In the last chapter of this dissertation, I will underline the main claims deriving from this study. The aim is to pull together the key findings and discuss their implications in the school world. I will also reflect on the role of research findings in the everyday school life.

It has not been an easy task to conduct research on a topic on which practically everybody has an opinion, as everyone has had some experiences of school, at least as a former student, or as a parent. Furthermore, policymakers have their own expectations in order to find an optimal balance between economic and school effectiveness factors. It is also clear that teachers interpret the research findings through the lens of their prior experience and understanding, and thus, translate the findings in the context of their everyday activities.

It is in the nature of scholarly research that the findings of a study must be interpreted and understood within the limits set by the data, methods and the overall framework of the study. Research often also calls for more investigation. However, the "on the one hand—on the other hand'' approach can be rather unhelpful for policymakers and teaching personnel (Harker & Tymms, 2004). As the limitations and methodological reflections have been presented above in this chapter, it is worthwhile to provide some implications and conclusions for the reader.

The theoretical importance of the possible existence of class compositional effects is closely linked to the question whether *class matters*. With this study, I have argued that students are profoundly influenced by the classes they have been assigned to. It also confirmed the findings from the previous Finnish studies (e.g., Kupiainen & Hienonen, 2016; Thuneberg et al., 2015) that there are notable differences between classes. This notion raises questions. Do different practices of allocating students to classes lead to inequitable opportunities for students assigned to different classes? What aspects are comparable across different classes, and what aspects differ? To what extent does the differential allocation account for inequality of achievement among students assigned to different classes? The present study cannot provide definitive answers to these questions. However, some conclusions can be made.

In Study I, the focus was on class size which failed to prove its power as such. However, it does not mean that class size has no effect on classroom processes. The number of students in the class necessarily affects what a teacher has to deal with, and what a teacher can do (Blatchford, Baines et al., 2001). As Pedder (2006) in his review on class size research suggests, with various variables involved in the teaching-learning process it is difficult to be certain what findings are attributable to class size alone, rather than to the cumulative effect of other variables in classroom processes. Based on the findings from this study, it is safe to say that future research should aim to identify the mechanisms by which the class size and class compositional effects come about.

There is no tracking, streaming or ability grouping in Finland—at least not officially. Yet, the results from Study II implied a hidden tracking system. Students without SEN but placed in classes with students with SEN performed less well than their peers in classes without any students with SEN. This lower level of performance was seen from the beginning of the lower secondary education right after the students had been assigned to new classes at the seventh grade. These hidden tracking practices have been identified in other recent Finnish studies (Kupiainen & Hienonen, in preparation; Koivuhovi et al., 2016).

When schools use different intentional student allocation practices, status hierarchies may be created (Gamoran et al., 1995). At the other end of the class distribution are larger, well-performing emphasized classes with aptitude tests, and on the other end, small special classes consisting only of students with SEN and, in some cases, much lower academic expectations (Kupiainen & Hienonen, 2016; Kupiainen & Hotulainen, 2019). This can also lead to a situation in which the classes between these two ends collect the remaining students and in the end, this may result in a widening of the achievement gap between high-level and low-level classes. One can argue that when both ends are ignored, the between-class differences would disappear. Yet, the differences between classes are the reality in Finnish education.

Furthermore, the performance differences often involve differences in family background and socioeconomic status, and thus, have unintended consequences (Gamoran et al., 1995). It is clear that the differences in student backgrounds within schools coincides with the differences in the wider society. The processes of allocating students can magnify achievement inequality, and thus, to reinforce the differences that students bring with them to school.

The aim to create homogenous classrooms is easily understandable and in line with the view to make teachers work more manageable (e.g., Hanushek & Wößmann, 2006; Kuzmina & Ivanova, 2017). However, there is little agreement on what constitutes the best teaching methods for different groups (Gamoran et al., 1995). When student allocation is done intentionally, the crucial question is, how to make most of it teaching-wise. Schools use their autonomy when they assign students into different classes, in terms of the size of the class and the composition of the class. Even though there might be a discrepancy between policy-level ideology and actual school practice, it is not likely that one would like to restrict this school autonomy and control the practices behind allocating students to classes. However, student allocation has real-world consequences and meaning for students. What must be understood better is the extent to which student placement practices are effectively, if not formally, random; whether student achievement matters as much as assumed when students are assigned to classrooms; how students are otherwise assigned to classrooms. Administrators, principals and teachers need to be aware of the grounds on which the placement

decisions are made and that the different placement practices have different consequences.

Evidence from Study II suggested that students with SEN placed in regular class benefited from other students with SEN in class. At the same time, partial evidence suggested that students without SEN, placed with students with SEN lost ground when the share of students with SEN increased. It seems that it is important in heterogeneous classes that students have classmates with a similar performance level and also classmates that perform slightly better (Kuzmina & Ivanova, 2017). Furthermore, the findings revealed that the more students with SEN there where in class, the lower was the performance level of other students. In addition, Study II also revealed that when the proportion of students with SEN in class exceeded a certain point, the performance level of the other students declined to the same level of students with SEN. It is a tempting idea to set a limit on the number of students with SEN in regular class. However, the process is much more complex, and it is bound with various variables in classroom processes. What can be said is that more attention should be paid to ensure that no one is placed in a needlessly restrictive environment in classes in which the target level of performance is reduced (Goldenberg, 1992).

Study III focused on the students at Tier 3 level and their class placement. Students with SEN placed in regular or special classes did not differ in any cognitive tasks. However, students in special classes received higher grades in some of the core subjects. These findings do not suggest that the placement in special class result in better academic achievement or higher grades, but they suggest that placement decisions can have unintended consequences. If teachers adjust their grades to the overall competence level of the class, corresponding performance in special classes may produce higher grades in special classes than in regular classes (see also Ouakrim-Soivio, 2013). In terms of the learning motivation, results implied a slightly different goal orientation structure in special classes—stronger aspiration for aiming good grades and outperforming classmates. Thus, these findings provide evidence that the reference group is associated with students' learning orientation. However, to draw any conclusions on the cause and effect, greater understanding in the form of future research is needed on the extent to which the non-cognitive outcomes of students with SEN relate to the educational performance and how this relationship differs across contexts.

In general, placement of students with SEN includes many issues: political, philosophical, social, physical, pedagogical, and emotional. The present study focused on cognitive aspects and took a brief look on the goal orientation. In terms of inclusive education, the dominant way to conceptualize it is to contrast special education provision in small special classes with the support provided in regular classes (Honkasilta et al., 2019), and rather than see them as a continuum, they are seen as dichotomies. However, being placed in a regular classroom with

support needs does not necessarily guarantee inclusive settings nor does the placement in special class mean inevitably segregated settings. A recognition of the importance of the provided support no matter the placement is a key to a successful learning. In some cases, this might require the recognition of the need for some students to be educated in smaller special classes (Kauffman, Nelson et al., 2017).

Kauffman and his various colleagues have been warning about 'place over the instruction' thinking, and stressed that a place can only be as good as the instruction students receive there (Kauffman & Badar 2014; Kauffman, Nelson et al. 2017; Kauffman & Pullen, 1996). They continue arguing that rather than considering the location or place of education, we should consider what kind of instruction and support services are necessary to optimize students' learning (Kauffman, Nelson et al. 2017). Furthermore, in theory, effective instructional approaches can be provided in both settings (Deno, 1970). However, this might be a too simplistic argument. Successful education of students with SEN is affected by the everyday classroom contexts (Blatchford & Webster, 2018), and this study lent support to these claims.

To provide the best possible education and guarantee better educational outcomes for every student is an honorable aim. In an ideal world, the well-known quote with respect to assessment: "Narrow the gap to raise the bar" (Hargreaves & Braun, 2013) also applies to assigning students to classes. With a careful student allocation, it could be ensured that every student can get appropriate instruction and adequate support. At the same time, the target level for every student would be set high enough yet achievable. However, some trade-offs maybe required. In real school life, it may mean supporting one group versus another. If in regular class we focus on students with SEN, what happens to the other students placed in these classes? To what extent should achievement gains of high-performing students be neglected for gains of low-performing or average students?

In this study I did not take a stand on how each school should allocate their students, however, the act of student allocation is not neutral. Trying to understand and explain the appearance of the compositional effect is therefore crucial.

Even though assigning students to a classroom plays an important role, there is more to it. Both teaching and peer interaction occur once the classroom door is closed and not only by organizing which students are behind those doors. In other words, not all can be managed by these processes. Future research calls for a researcher to step into a classroom. Detailed observational data about instructional processes and student-teacher interactions could unveil some of the mechanisms mediating the class placement and class composition effects.

Going back to the beginning, I set the research tasks for this study after I had discovered a need for a study on placement of students with SEN in the Finnish context. Along the way, there were moments of doubt. However, in the end, this study managed to fulfill its task, test all the hypotheses and answer all the research

questions. They may not be the definitive answers, but I hope they will serve as a foundation for future research. I also hope that this study will add to understanding the effects of student placement and classroom assignment processes. Finally, I hope that these results will facilitate the discussion of class placement of students with SEN by providing evidence.

References

Akerhielm, K. (1995). Does class size matter? *Economics of Education, 14*, 229–241.

Alcott, L. M. (1871). *Little Men.* New York, NY: Puffin.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93–114.

Ames, C. (1992). Classrooms: Goals, Structures, and Student Motivation. *Journal of Educational Psychology, 84*(3), 261–271.

Bakker, J. T. A., &. Bosman, A. M. T. (2003). Self-Image and Peer Acceptance of Dutch Students in Regular and Special Education. *Learning Disability Quarterly, 26*(1), 5–14. https://doi.org/10.2307/1593680

Bakker, J., Denessen, E., Bosman, A., Krijger, E.-M., & Bouts, L. (2007). Sociometric status and self-image of children with specific and general learning disabilities in Dutch general and special education classes. *Learning Disability Quarterly, 30*, 47–62. https://doi.org/10.2307/30035515

Basic Education Act 628/1998. Amendments up to 1136/2010. Government of Finland. Retrieved [15.04.2019] from http://www.finlex.fi/en/laki/kaannokset/1998/en19980628.pdf/.

Basic Education Decree 852/1998. Amendments up to 966/2016. Government of Finland. Retrieved [15.04.2019] from http://www.finlex.fi/fi/laki/ajantasa/1998/199808527

Bear, G. G., Minke, K. M., & Manning, M. A. (2002). Self-concept of students with learning disabilities: A meta-analysis. *School Psychology Review, 31,* 405–427.

Becker S., & Ichino A. (2002). Estimation of average treatment effects based on propensity scores. *Stata Journal, 2(4)*, 358–377.

Belfi, B., Goos, M., De Fraine, B., & Van Damme, J. (2012). The effect of class composition by gender and ability on secondary school students' school well-being and academic self-concept: A literature review. *Educational Research Review, 7,* 62–74. https://doi.org/10.1016/j.edurev.2011.09.002

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researchers, 31*(8), 18–20.

Betts, J. R., & Shkolnik, J. L. (1999). The behavioral effects of variations in class size: The case of math teachers. *Educational Evaluation and Policy Analysis, 21*(2), 193–213.

Björn, P., Aro, M. T., Koponen, T. K., Fuchs, L. S., & Fuchs, D. H. (2015). The many faces of special education within RTI frameworks in the United States and Finland. *Learning Disability Quarterly, 39*(1), 58–66. https://doi:10.1177/073194871559478

Bland, J. M., & Altman, D. G. (1994). Some examples of regression towards the mean. *BMJ, 24,* 309. https://doi.org/10.1136/bmj.309.6957.780

Blatchford, P., Baines, E., Kutnick, P., & Martin, C. (2001). Classroom contexts: Connections between class size and within class grouping. *British Journal of Educational Psychology, 71*, 283–302. https://doi:10.1348/000709901158523

Blatchford, P., Bassett, P., & Brown, P. (2005). Teachers' and Pupils' Behavior in Large and Small Classes: A Systematic Observation Study of Pupils Aged 10 and 11 Years. *Journal of Educational Psychology, 97*(3), 454–467. https://doi:10.1037/0022-0663.97.3.454

Blatchford, P., Basset, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher-pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and Instruction, 21*, 715–730. https://doi:10.1016/j.learninstruc.2011.04.001

Blatchford, P., Basset, P., Goldstein, H., & Martin, C. (2003). Are class size differences related to pupils' educational progress and classroom processes? Findings from the institute of education class size study of children aged 5–7 years. *British Educational Research Journal, 29*, 709–730. DOI:10.1080/0141192032000133668

Blatchford, P., Edmonds, S., & Martin, C. (2003). Class size, pupil attentiveness and peer relations. *British Journal of Educational Psychology, 73*, 15–36. https://doi.org/10.1348/000709903762869897

Blatchford, P., Goldstein, H., Martin, C., & Browne, W. (2002). A Study of Class Size Effects in English School Reception Year Classes. *British Educational Research Journal, 28*(2), 169–185. https://doi.10.1080 /0141192012012213 0

Blatchford, P., Kutnick, P., Baines E., & Galton, M. (2003). Toward a social pedagogy of classroom group work. *International Journal of Educational Research 39,* 153–172. https://doi:10.1016/S0883-0355(03)00078-8

Blatchford, P., & Martin. C. (1998). The Effects of Class Size on Classroom Processes: 'It's a Bit Like a Treadmill – Working Hard and Getting Nowhere Fast!*' British Journal of Educational Studies, 46*(2), 118–137. https://doi.org/10.1111/1467-8527.00074

Blatchford, P., Moriarty, V., Edmonds, S., & Martin, C. (2002). Relationships Between Class Size and Teaching: A Multimethod Analysis of English Infant Schools. *American Educational Research Journal, 39*(1), 101–132. https://doi:10.3102/00028312039001101

Blatchford, P., & Mortimore, P. (1994). The issue of class size in schools: What can we learn from research*? Oxford Review of Education, 20*(4), 411–428. https://doi.10.1080/0305498940200402

Blatchford, P., & Russell, A. (2019). New ways of thinking about research on class size: An international perspective. Introduction to the special section. *International Journal of Educational Research, 96*, 120–124. https://doi.org/10.1016/j.ijer.2018.09.011

Blatchford, P., & Webster, R. (2018). Classroom contexts for learning at primary and secondary school: Class size, groupings, interactions and special educational needs. *British Educational Research Journal, 44*(4), 681–703. https://doi.org/10.1002/berj.3454

Botsas, G., & Padeliadu, S. (2003). Goal orientation and reading comprehension strategy use among students with and without reading difficulties. *International Journal of Educational Research 39*, 477–495. https://doi:10.1016/j.ijer.2004.06.010

Bourke, S. (1986). How Smaller Is Better: Some Relationships Between Class Size, Teaching Practices, and Student Achievement. *American Educational Research Journal, 23*(4), 558–571.

Bryk A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Advanced Quantitative Techniques in the Social Sciences Series 1. Newbury Park, CA: Sage Publications.

Byrne, B. M. (2011). *Structural Equation Modeling with Mplus. Basic Concepts, Applications, and Programming.* New York, NY: Routledge.

Caliendo, M., & Kopeing, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22,* 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.xS

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally & Company.

Carlberg, C., & Kavale, K. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. *Journal of Special Education, 14*(3), 265–309.

Cheng, R. W., McInerney, D. M., & Mok, M. C. C. (2014). Does big-fish–little-pond effect always exist? Investigation of goal orientations as moderators in the Hong Kong context, Educational Psychology: *An International Journal of Experimental Educational Psychology, 34*(5), 561–580. https://doi:10.1080/01443410.2014.898740

Cheung, C. W., & Lau, R. S. (2008). Testing Mediation and Suppression Effects of Latent Variables Bootstrapping With Structural Equation Models. *Organizational Research Methods, 11*(2), 296–325. https://doi.org/10.1177/1094428107300343

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, C., Waldron, N., & Majd, M. (2004). Academic progress of students across inclusive and traditional settings. *Mental Retardation, 42*, 136–44. https://doi:10.1352/0047-6765(2004)42<136:APOSAI>2.0.CO;2

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1996). *Equality of educational opportunity*. Washington, DC: US Department of Health, Education & Welfare. Office of Education.

Crick, R., Ren, K., & Stringher, C. (2014). Introduction. In R. Crick, C. Stringer, & K. Ren (Eds.), *Learning to learn. International perspectives from theory and practice* (pp. 1–6). London: Routledge.

Crick, R., Stringer, C., & K. Ren. (Eds.) (2014). Learning to learn. International perspectives from theory and practice. London: Routledge.

Cuong, N. V. (2013). Which covariates should be controlled in propensity score matching? Evidence from a simulation study. *Statistica Neerlandica*, *67*(2), 169–180. https://doi:10.1111/stan.12000

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guilford Press.

Demetriou, A. (2014). Learning to learn, know and reason. In R. Crick, C. Stringer, & K. Ren (Eds.), *Learning to learn. International perspectives from theory and practice* (pp. 41–66). London: Routledge.

Demetriou, A., Pachaury, A., Metallidou, Y., & Kazi, S. (1996). Universals and specificities in the structure and development of quantitative-relational thought: A cross-cultural study in Greece and India. *International Journal of Behavioural Development, 19*(2), 255–290. https://doi:10.1080/016502596385785

Deno, E. (1970). Special education as developmental capital. *Exceptional Children, 37*(3), 229–237.

Dessemontet, R. S., Bless, G., & Morin, D. (2012). Effects of inclusion on the academic achievement and adaptive behaviour of children with intellectual disabilities. *Journal of Intellectual Disability Research, 56,* 579–587. https://doi.10.1111/j.1365-2788.2011.01497.x

Dobbelsteen, S., Levin, J., & Oosterbeek, H. (2002). The causal effect of class size on scholastic achievement: distinguishing the pure class size effect from the effect of changes in class composition. *Oxford Bulletin of Economics and Statistics, 64*(17), 17–38. https://doi.org/10.1111/1468-0084.00003

Dockx, J., De Fraine, B., & Vandecandelaere, M. (2019). Tracks as frames of reference for academic self-concept. *Journal of School Psychology, 72*, 67–90. https://doi.org/10.1016/j.jsp.2018.12.006

Dreeben, R., & Barr, R. (1988). Classroom Composition and the Design of Instruction. *Sociology of Education, 61*(3), 129–142.

Dweck, C. S. (1986). Motivational Processes Affecting Learning. *American Psychologist, 41*(10), 1040–1048.

EASIE. (2016). European Agency Statistics on Inclusive Education (EASIE) Methodology Report. (A. Watkins, S. Ebersold and A. Lénárt, eds.). Odense, Denmark.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest, 2*(1), 1–30. https://doi.org/10.1111/1529-1006.003

Elbaum, B. (2002). The Self-Concept of Students with Learning Disabilities: A Meta-Analysis. *Learning Disabilities Research & Practice, 17*(4), 216–226. https://doi.org/10.1111/1540-5826.00047

Eliot, T. S. (1933). *Selected essays*. London: Faber and Faber Limited.

Elliot, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*, 5–12.

Elliott, J., Thurlow, M., & Ysseldyke, J. (1996). *Assessment guidelines that maximize the participation of students with disabilities in large-scale assessments: Characteristics and considerations*. Retrieved [15.10.2019] from www.cehd.umn.edu/NCEO/onlinepubs/archive/Synthesis/synthesis25.html

European Commission. (2019). *Access to quality education for children with special educational needs.* Retrieved [29.10.2019] from https://op.europa.eu/en/publication-detail/-/publication/b2215e85-1ec6-11e9-8d04-01aa75ed71a1/language-en

European Union. (2018). *Council Recommendation on promoting common values, inclusive education, and the European dimension of teaching.*

Retrieved [29.9.2019] from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018H0607(01)

Farrell, P., Dyson, A., Polat, F., Hutcheson, G., & Gallannaugh, F. (2007). SEN Inclusion and Pupil Achievement in English Schools. *Journal of Research in Special Educational Needs, 7*(3), 172–178. https://doi.org/10.1111/j.1471-3802.2007.00094.x

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A Statewide Experiment. *American Educational Research Journal, 27*(3), 557–77.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's Class Size Study: Findings, Implications, Misconceptions. *Educational Evaluation and Policy Analysis, 21*(2), 97–109.

Finnish Advisory Board on Research Integrity. (2012). *Responsible conduct of research and procedures for handling allegations of misconduct in Finland*. Helsinki: Finnish Advisory Board on Research Integrity.

Fletcher, J. (2009). The effects of inclusion on classmates of students with special needs: The case of serious emotional problems. *Education Finance and Policy, 4,* 278–299.

Fletcher, J. (2010). Spillover effects of inclusion of classmates with emotional problems on test scores in early elementary school. *Journal of Policy Analysis and Management, 29*(1), 69–83. https://doi.org/10.1002/pam.20479

Florian, L. (2014). Inclusive pedagogy. An alternative approach to difference and inclusion. In F. Kiuppis & R. Sarromaa Hausstätter (Eds.), *Inclusive education. Twenty years after Salamanca* (pp. 219–229). New York, NY: Peter Lang.

FNBE. (2004). *National Core Curriculum for Basic Education*. Helsinki: National Board of Education.

FNBE. (2016). *National Core Curriculum for Basic Education.* Publications 2016:5. Helsinki: National Board of Education.

FNBE. (2019). *Vastauksia tuen kysymyksiin*. Retrieved [15.11.2019] from https://www.oph.fi/fi/koulutus-ja-tutkinnot/vastauksia-tuen-kysymyksiin

Fore, C., Hagan-Burke, S., Burke, M., Boon, R., & Smith, S. (2008). Academic achievement and class placement in high school: Do students with learning disabilities achieve more in one class placement than another? *Education and Treatment of Children, 31*, 55–72. https://doi:10.1353/etc.0.0018

Fuchs, D., Fuchs, L. S., McMaster, K. L., & Lemons, C. J. (2018). Students with Disabilities' Abysmal School Performance: An Introduction to the Special Issue. *Learning Disabilities Research & Practice, 33*(3), 127–130. https://doi:10.1111/ldrp.12180

Gamoran, A., Nystrand, M., Berends, M., & LePore, P. C. (1995). An Organizational Analysis of the Effects of Ability Grouping. *American Educational Research Journal, 32*, 687–715.

Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing High-Quality Research in Special Education: Group Experimental Design. *Journal of Special Education, 34*, 2–15. https://doi.10.1177/002246690003400101

Gersten, R., Fuchs, L., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. (2005). Quality Indicators for Group Experimental and Quasi-Experimental Research in Special Education. *Exceptional Children, 71*(2), 149–164. https://doi.org/10.1177/001440290507100202

Gersten, R., Jaynthi, M., Santoro, L., & Newman-Conchar, R. (2017). Designing Rigorous Group Studies in Special Education. In J. M. Kauffman, D. P. Hallahan & P. Cullen Pullen (Eds.), *Handbook of Special Education* (pp. 105–115). New York, NY: Education Routledge.

Gersten, R., & Smith-Johnson, J. (2001). Reflections on the Research to Practice. *Teacher Education and Special Education, 24*, 356–361. https://doi.org/10.1177/088840640102400409

Glass, G. V., & Smith, M. L. (1979). Meta-Analysis of Research on Class Size and Achievement. *Educational Evaluation and Policy Analysis, 1*(1), 2–16. https://doi.org/10.3102/01623737001001002

Gnambs, T., & Nusser, L. (2019). The Longitudinal Measurement of Reasoning Abilities in Students With Special Educational Needs. *Frontiers in Psychology, 10*(232). https://doi.org/10.3389/fpsyg.2019.00232

Goldenberg, J. (1992). The Limits of Expectations: A case for case knowledge about teacher expectancy effects. *American Educational Research Journal, 29*, 517–544. https://doi.org/10.3102/00028312029003517

Gorges, J., Neumann, P., Wild, E., Stranghöner, D., & Lütje-Klose, B. (2018). Reciprocal effects between self-concept of ability and performance: A longitudinal study of children with learning disabilities in inclusive versus exclusive elementary education. *Learning and Individual Differences, 61*, 11–20. https://doi.org/10.1016/j.lindif.2017.11.005

Graham, L. J., & Jahnukainen, M. (2011). Where art thou, inclusion? Analysing the development of inclusive education in New South Wales, Alberta and Finland. *Journal of Education Policy, 26*(2), 263–288. https://doi:10.1080/02680939.2010.493230

Greenland, S. (2017). For and against methodologies: Some perspectives on recent causal and statistical inference debates. *European Journal of Epidemiology, 23*(1), 3–20. https://doi.org/10.1007/s10654-017-0230-6.

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis, 21*(2), 143–163.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The Economic Journal, 113,* F64–F98.

Hanushek, E. A., Kain, J., & Rivkin, S. (2002). Inferring program effects for special populations: Does special education raise achievement for students with disabilities? *Review of Economics and Statistics, 84,* 584–599. https://doi.org/10.1162/003465302760556431

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal, 116*(510), C63–C76. https://doi.org/10.1111/j.1468-0297.2006.01076.x

Harfitt, G., & Tsui, A. (2015). An examination of class size reduction on teaching and learning processes: A theoretical perspective. *British Educational Research Journal, 41*, 845–865. https://doi:10.1002/berj.3165

Hargreaves, A., & Braun, H. (2013). *Data-Driven Improvement and Accountability*. Boulder, CO: National Education Policy Center. Retrieved [22.2.2020] from http://nepc.colorado.edu/publication/data-driven-improvement-accountability/

Hargreaves, L., Galton, M., & Pell, A. (1998). The effects of changes in class size on teacher-pupil interactions. *International Journal of Educational Research, 29*, 779–795.

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement, 15*(2), 177–199.

Hattie, J. (2002). Classroom composition and peer effects. *International Journal of Educational Research 37*, 449–481. http://doi:10.1016/S0883-0355(03)00015-6

Hattie, J. (2005). The paradox of reducing class size and improved learning outcomes. *International Journal of Educational Research, 42*, 387–425. http://doi:10.1016/j.ijer.2006.07.002

Hattie, J. (2009). *Visible learning. A synthesis of meta-analyses relating to achievement.* New York, NY: Routledge.

Hautamäki, J. (1984). *Peruskoululaisten loogisen ajattelun mittaamisesta ja esiintymisestä.* Joensuun yliopiston yhteiskuntatieteellisiä julkaisuja 1. Joensuu: Joensuun yliopisto.

Hautamäki, J., Arinen, P., Niemivirta, M., Eronen, S., Hautamäki, A., Kupiainen, S., Lindblom, B., Pakaslahti, L., Rantanen, P., & Scheinin, P. (2002). *Assessing Learning-to-learn: A Framework.* Evaluation 4/2002. Helsinki: National Board of Education.

Hautamäki, J., & Kupiainen, S. (2014). Learning to Learn in Finland. In R. Crick, C. Stringer & K. Ren (Eds.), *Learning to Learn: International Perspectives from Theory and Practice* (pp. 170–195). London: Routledge.

Hautamäki, J., Kupiainen, S., Marjanen, J., Vainikainen, M.-P., & Hotulainen, R. (2013). *Oppimaan oppiminen peruskoulun päättövaiheessa: Tilanne vuonna 2012 ja muutos vuodesta 2001.* Tutkimuksia 347. Helsinki: Helsingin yliopisto.

Heck, R., H., Thomas, S. L., & Tabata L. N. (2010). *Multilevel and Longitudinal Modeling with IBM SPSS*. New York, NY: Routledge.

Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013). Including students with special educational needs into large-scale assessments of competencies: challenges and approaches within the German National Educational Panel Study (NEPS). *Journal for educational research online, 5*(2), 217–240.

Hibel, J., Farkas, G., & Morgan, P. (2010). Who Is Placed into Special Education? *Sociology of Education, 83*, 312–332. http://doi:10.1177/0038040710383518

Hienonen, N., & Lintuvuori, M. (2018). Opetuksen toteutuspaikka yläkoulussa – erilaiset opetusryhmät ja osaaminen. In *Oppimisen tuki*

varhaislapsuudesta toisen asteen siirtymään: tasa-arvon toteutuminen ja kehittämistarpeet (pp. 75–81). Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 55/2018. Helsinki: Valtioneuvoston kanslia.

Holmes, W. M. (2014). *Using Propensity Scores in Quasi-Experimental Designs*. California, CA: Sage Publications.

Hong, G. (2015). *Causality in a social world. Moderation, mediation and spill-over*. Chichester: John Wiley & Sons.

Honkasilta, J., Ahtiainen, R., Hienonen, N., & Jahnukainen, M. (2019). Inclusive and Special Education and the Question of Equity in Education: The Case of Finland. In M. J. Schuelka, C. J. Johnstone, G. Thomas, & A. J. Artiles (Eds.), *The SAGE Handbook on Inclusion and Diversity in Education* (pp. 481–495). London: Sage Publications.

Hosenfeld, B., van den Boom, D. C., & Resing, W .C .M. (1997). Constructing geometric analogies test for the longitudinal testing of elementary school children. *Journal of Educational Measurement, 34*(4), 367–372. https://doi.org/10.1111/j.1745-3984.1997.tb00524.x

Hoskins, B., & Fredriksson, U. (2014). *Learning to Learn: What is it and can it be measured?* Luxembourg: Office for Official Publications of the European Communities.

Hox, J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hox, J. J., & Bechger, T. M. (1998). Introduction to structural equation modeling. *Family Science Review, 11,* 354–373.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics, 115*, 1239–1285. https://doi.org/10.1162/003355300555060

Hoxby, C. M., & Weingarth. G. (2005). Taking race out of the equation: School reassignment and the structure of peer effects. Manuscript, Department of Economics, Harvard University. Retrieved [17.10.2019] from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.75.4661

Huber, K. D., Rosenfeld, J. G., & Fiorello, G. A. (2001). The differential impact of inclusion and inclusive practices on high, average and low achieving general education students. *Psychology in the Schools, 38*(6), 497–504. https://doi.org/10.1002/pits.1038

Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Seaton, M., Nezlek, J., Suls, J., & Régner, I. (2009). Clarifying the he role of social comparison in the big-fish–little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology, 97*(1), 156–170. https://doi.org/10.1037/a0015558

Huguet, P., Dumas, F., Monteil, J. M., & Genestoux, N. (2011). Social comparison choices in the classroom: further evidence for students' upward comparison tendency and its beneficial impact on performance. *European Journal of Social Psychology, 31,* 557–578. https://doi.org/10.1002/ejsp.81

Insel P. M., & Lindgren, H. C. (1978). *Too Close for Comfort: The Psychology of Crowding.* University of Michigan: Prentice-Hall.

Jahnukainen, M. (2015). Inclusion, integration, or what? A comparative study of the school principals' perceptions of inclusive and special education in Finland and in Alberta, Canada. *Disability & Society, 30*(1)*,* 59–72. https://doi.org/10.1080/09687599.2014.982788

Jahnukainen, M., & Itkonen, T. (2015). Tiered intervention: History and trends in Finland and the United States. *European Journal of Special Needs Education, 31*(1), 140–150. https://doi.org/10.1080/08856257.2015.1108042

Juster, N. (1961). The Phantom Tollbooth. Epstein & Carroll.

Juva, S. (2008). Inhimillinen pääoma ja koulutuksen tehokkuus – koulutus taloustieteen tutkimuskohteena. In J. Heikkilä, S. Juva, T. Kettunen, M. Lahtinen, & R. Tiihonen (Eds.), *Koulutuksen talouden käsikirja* (pp. 15–26). Jyväskylä: PS-kustannus.

Kalambouka, A., Farrell, P., Dyson A., & Kaplan, I. (2007). The impact of placing pupils with special educational needs in mainstream schools on the achievement of their peers. *Educational Research, 49,* 365–382. https://doi.org/10.1080/00131880701717222

Karjalainen, T., & Lamberg, K. (2017) Esi- ja perusopetuksen opetusryhmät. In T. Kumpulainen (Ed.), *Opettajat ja rehtorit suomessa 2016* (pp. 195–207). Raportit ja selvitykset 2017:2. Helsinki: Opetushallitus.

Kauffman, J. M., & Badar, J. (2014). It's instruction over place—not the other way around! *Phi Delta Kappan, 98*(4), 55–59. https://doi.org/10.1177/0031721716681778

Kauffman, J. M., & Lloyd, J. W. (2017). Statistics, Data, and Special Educational Decisions. In J. M. Kauffman, D. P. Hallahan, & P. Cullen Pullen (Eds.), *Handbook of Special Education* (pp. 29–39). New York, NY: Education Routledge.

Kauffman, J. M., Nelson, C. M., Simpson, R. L., & Ward, D. M. (2017). Contemporary issues. In J. M. Kauffman, D. P. Hallahan, & P. Cullen Pullen (Eds.), *Handbook of Special Education* (pp. 16–28). New York, NY: Education Routledge.

Kauffman, J. M., & Pullen, P. L. (1996). Eight myths about special education. *Focus on Exceptional Children, 28*(5), 1–12. DOI: 10.17161/foec.v28i5.6854

Keslair, F., Maurin, E., & McNally, S. (2012). Every child matters? An evaluation of "Special Educational Needs" programmes in England. *Economics of Education Review, 31*(6)*,* 932–948. https://doi.org/10.1016/j.econedurev.2012.06.005

Kintsch, W. & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review, 8*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kiuppis, F., & Sarromaa Hausstätter, R. (2014). Inclusive education for all, and especially for someone. In F. Kiuppis & R. Sarromaa Hausstätter (Eds.), *Inclusive education. Twenty years after Salamanca* (pp. 1–5). New York, NY: Peter Lang.

Kivinen, U. (2009). Erityiskasvatus ammattialana. In S. Moberg, J. Hautamäki, J. Kivirauma, U. Lahtinen, H. Savolainen & S. Vehmas (Eds.), *Erityispedagogiikan perusteet* (pp. 171–194). Jyväskylä: PS-kustannus.

Kivirauma, J., & Ruoho, K. (2007). Excellence through Special Education? Lessons from the Finnish School Reform. *International Review of Education, 53*(3), 283–302. DOI: 10.1007/s11159-007-9044-1

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling* (2nd ed.). New York: The Guildford Press.

Koivuhovi, S., Vainikainen, M.-P., Kalalahti M. & Niemivirta, M. (2019). Changes in Children's Agency Beliefs and Control Expectancy in Classes With and Without a Special Emphasis in Finland from Grade Four to Grade Six. *Scandinavian Journal of Educational Research, 63*(3), 427–442. https://doi.org/10.1080/00313831.2017.1402364

Kojac, A., Kuhl, P., Jansen, M., Pant, H. A., & Stanat, P. (2018). Educational placement and achievement motivation of students with special educational needs. *Contemporary Educational Psychology, 55,* 63–83. https://doi.org/10.1016/j.cedpsych.2018.09.004

Konstantopoulos, S., & Traynor, A. (2014). Class Size Effects on Reading Achievement Using PIRLS Data: Evidence from Greece. *Teachers College Record, 116*(2).

Kosunen, S. (2016). *Families and the social space of school choice in urban Finland.* Studies in Educational Sciences 267. Helsinki: University of Helsinki.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114*(2), 497–532. https://doi.org/10.1162/003355399556052

Kupiainen, S., & Hienonen, N. (2016). *Luokkakoko.* Kasvatusalan tutkimuksia 72 Jyväskylä: Suomen kasvatustieteellinen seura.

Kupiainen, S., & Hotulainen, R. (2019). Erilaisia luokkia, erilaisia oppilaita. In J. Hautamäki, I. Rämä, & M.-P. Vainikainen (Eds.), *Perusopetus, tasa-arvo ja oppimaan oppiminen: Valtakunnallinen arviointitutkimus peruskoulun päättövaiheesta* (pp. 139–165). Kasvatustieteellisiä tutkimuksia 52. Helsinki: Helsingin yliopisto.

Kupiainen, S., Vainikainen, M.-P., Marjanen, J., & Hautamäki, J. (2014). The role of time on task in computer-based low-stakes assessment of cross-curricular skills. *Journal of Educational Psychology, 106*(3), 627–638. https://doi.org/10.1037/a0035507

Kuzmina, Y., & Ivanova, A. (2017). The effects of academic class composition on academic progress in elementary school for students with different levels of initial academic abilities. *Learning and Individual Differences, 64*, 43–53. https://doi.org/10.1016/j.lindif.2018.04.004

Lahtinen, M., & Lankinen T. (2015). *Koulutuksen lainsäädäntö käytännössä.* Helsinki: Tietosanoma Oy.

Lane, F., Yen, M., Shelley, K., & Henson, R. (2012). An illustrative example of propensity score matching with education research. *Career and Technical Education Research, 37*(3), 187–212. https://doi.org/10.5328/cter37.3.187

Lee, V. E. (2000). Using Hierarchical Linear Modeling to Study Social Contexts: The Case of School Effects. *Educational Psychologist, 35*(2), 125–141. https://doi.org/10.1207/S15326985EP3502_6

Lei, P., & Wu, Q. (2007). Introduction to structural equation modeling: issues and practical considerations. *Educational Measurement: Issues and Practice, 26*(3), 33–43. https://doi.org/10.1111/j.1745-3992.2007.00099.x

Leino, K., Ahonen, A. Hienonen, N., Hiltunen, J., Lintuvuori, M., Lähteinen, S., Lämsä, J., Nissinen, K., Nissinen, V., Puhakka, E., Pulkkinen, J., Rautopuro, J., Sirén, M., Vainikainen, M.-P., & Vettenranta, J. (2019). *PISA18 Ensituloksia. Suomi parhaiden joukossa. Suomalaisnuorten lukutaidon ja lukuharrastuksen muuttuminen*. Opetus- ja kulttuuriministeriön julkaisuja 2019:40. Helsinki: Opetus- ja kulttuuriministeriö.

Lindsay, G. (2003). Inclusive education: a critical perspective. *British Journal of Special education, 30*(1), 3–12. https://doi.org/10.1111/1467-8527.00275

Lintuvuori, M. (2019). *Perusopetuksen oppimisen ja koulunkäynnin tuen järjestelmän kehitys tilastojen ja normien kuvaamana.* Kasvatustieteellisiä tutkimuksia 51. Helsinki: Helsingin yliopisto.

Lintuvuori, M., Hienonen, N., & Hautamäki, J. (2019). Oppimaan oppimisen arviointi tehostetun ja erityisen tuen näkökulmasta. In J. Hautamäki, I. Rämä, & M.-P. Vainikainen (Eds.), *Perusopetus, tasa-arvo ja oppimaan oppiminen: Valtakunnallinen arviointitutkimus peruskoulun päättövaiheesta* (pp. 125–137). Kasvatustieteellisiä 52. Helsinki: Helsingin yliopisto.

Luellen, J., Shadish, W. R., & Clark, M. H. (2005). Propensity scores. An Introduction and Experimental Test. *Evaluation Review, 29*, 530–558. https://doi.org/10.1177/0193841X05275596

Lüdtke, O., Robitzsch, A., Asparouhov. T., Marsh, H. W., Trautwein, U., & Muthén, B. (2008). The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies. *Psychological Methods, 13*(3), 203–229. https://doi.org/10.1037/a0012869

Lyytinen, S., & Lehto, J. E. (1998). Hierarchy rating as a measure of text macroprocessing: Relationship with working memory and school achievement. *Educational Psychology, 18*(2), 157–169.

Maas, C. J. M., & Hox, J. H. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal Multilevel Models of the Big Fish Little Pond Effect on Academic Self-concept: Counterbalancing Contrast and Reflected Glory Effects in Hong Kong Schools. *Journal of Personality and Social Psychology, 78*(2), 337–349. https://doi.org/10.1037/0022-3514.78.2.337

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106–124. https://doi.org/10.1080/00461520.2012.670488

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology, 47*(1), 213–231. https://doi.org/10.1037/0022-3514.47.1.213

May, K., & Hittner, J. B. (2010). Reliability and validity of gain scores considered graphically. *Perceptual and Motor Skills 111*(2), 399-406. https://doi.org/10.2466/03.PMS.111.5.399-406

McDonald, R. P., & Ho, M.-H. R (2002). Principles and Practice in Reporting Structural Equation Analyses. *Psychological Methods, 7*(1), 64–82. https://doi.org/10.1037/1082-989X.7.1.64

Ministry of Education and Culture. (2011). Opetusryhmien pienentämiseen haettavana avustusta. A press release 17.11.2011. Retrieved [15.11.2019] from http://www.minedu.fi/OPM/Tiedotteet/2011/11/ryhmxkokoavustus.html?lang=fi.

Ministry of Education and Culture. (2012). *83 miljoonaa euroa opetusryhmien pienentämiseen ja koulujen välisten erojen kaventamiseen.* A press release 3.1.2020. Retrieved [15.11.2019] from http://www.minedu.fi/OPM/Tiedotteet/2012/10/ryhmakoko_valtionavustus.html

Ministry of Education and Culture. (2017). *Perusopetuksen tasa-arvoa edistävät toimenpiteet: erityisopetuksen laadun kehittäminen ja siihen liittyvä opettajien ja koulunkäyntiavustajien palkkaaminen ja opetusryhmäkoon pienentäminen.* Retrieved [22.8.2019] from https://minedu.fi/avustukset/avustus/-/asset_publisher/perusopetuksen-tasa-arvoa-edistavat-toimenpiteet.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., & Ehrle, K. (1999). Evaluating the SAGE program: A pilot Program in targeted pupil-teacher reduction in Wisconsin. *Educational Evaluation and Policy analyses, 21*(2), 165–177.

Morgan, P., Frisco, M., Farkas, G., & Hibel, J. (2010). A Propensity Score Matching Analysis of the Effects of Special Education Services. *Journal of Special Education, 43,* 236–254. https://doi.org/10.1177/0022466908323007

Morningstar, M. E., Kurth, J. A., & Johnson, P. E. (2017). Examining National Trends in Educational Placements for Students with Significant Disabilities. *Remedial and Special Education, 38*, 3–12. https://doi.org/10.1177/0741932516678327

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017*). PIRLS 2016 International Results in Reading*. Retrieved [23.10.2019] from http://timssandpirls.bc.edu/pirls2016/international-results/wp-

content/uploads/structure/CompletePDF/P16-PIRLS-International-Results-in-Reading.pdf

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide version 7*. Retrieved [15.5.2019] from https://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r3_web.pdf

Myklebust, J. O. (2007). Diverging paths in upper secondary education: Competence attainment among students with special educational needs. *International Journal of Inclusive Education, 11,* 215–231. https://doi.org/10.1080/13603110500375432

National Advisory Board on Research Ethics. (2009). *The ethical principles of research with human participants and ethical review in the human sciences in Finland.* Finnish National Board on Research Integrity TENK guide. Tutkimuseettisen neuvottelukunnan julkaisuja 3/2019. Helsinki: Finnish National Board on Research Integrity TENK.

Niemi, H. (2015). Teacher Professional Development in Finland: Towards a More Holistic Approach. *Psychology, Society and Education, 7*(3), 278–294. https://doi.org 10.25115/psye.v7i3.519

Niemivirta, M. (2004). *Habits of mind and academic endeavors. The Correlates and Consequences of Achievement Goal Orientations*. Research Report 196. Helsinki: University of Helsinki.

Nye, B., Konstantopoulos, S., &, Hedges,L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257. https://doi.org/10.3102/01623737026003237

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in Special Education: Scientific Methods and Evidence-Based Practices. *Exceptional Children, 71*(2), 137–148. https://doi.org/10.1177/001440290507100201

OECD. (2007). *Students with Disabilities, Learning Difficulties and Disadvantages Policies, Statistics and Indicators*. Paris: OECD Publishing. http://www.oecd.org/education/school/40299703.pdf

OECD. (2012). *Equity and quality in education.* Paris: OECD Publishing. doi.org/10.1787/9789264130852-en.

OECD. (2013). *TALIS 2013 Results. An International Perspective on Teaching and Learning*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264196261-en

OEDC. (2014). *Education at a Glance 2014. OECD Indicators*. Paris: OECD Publishing. http://dx.doi.org/10.1787/eag-2014-en

OECD. (2016a). *Low-Performing Students: Why They Fall Behind and How to Help Them Succeed.* Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264250246-en.

OECD. (2016b). *PISA 2015 Results (Volume I): Excellence and Equity in Education.* Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264266490-en

OECD. (2017). *PISA 2015 Technical Report.* Paris: OECD Publishing.

OECD. (2019a). *Education at a Glance 2018. OECD Indicators*. Paris: OECD Publishing. https://doi.org/10.1787/19991487

OECD (2019b). *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners*. Paris: OECD Publishing. https://doi.org/10.1787/1d0bc92a-en

Oh-Young, C., & Filler, J. (2015). A meta-analysis of the effects of placement on academic and social skill outcome measures of students with disabilities. *Research in Developmental Disabilities, 47*, 80–92. https://doi.org/10.1016/j.ridd.2015.08.014

Opdenakker M.-C., & Van Damme, J. (2000) The Importance of Identifying Levels in Multilevel Analysis: An Illustration of the Effects of Ignoring the Top or Intermediate Levels in School Effectiveness Research. *School Effectiveness and School Improvement, 11*(1), 103–130. https://doi.org/10.1076/0924-3453(200003)11:1;1-A;FT103

OSF. (2019). *Special education.* Helsinki: Statistics Finland. Retrieved [23.6.2019] from http://www.stat.fi/til/erop/index.html

Ouakrim-Soivio, N. (2013). *Toimivatko päättöarvioinnin kriteerit? Oppilaiden saamat arvosanat ja Opetushallituksen oppimistulosten seuranta-arviointi koulujen välisten osaamiserojen mittareina.* Raportit ja selvitykset 2013:9. Helsinki: Opetushallitus.

Patrick, H., Kaplan, A., & Ryan, A. M. (2011). Positive classroom motivational environments: Convergence between mastery goal structure and classroom social climate. *Journal of Educational Psychology, 103*(2), 367–382. https://doi.org/10.1037/a0023311

Paufler, N. A., & Amrein-Beardsley, A. (2013). The random assignment of students into elementary classrooms: implications for value-added analyses and interpretations. *American Educational Research Journal, 51,* 328–362. https://doi.org/10.3102/0002831213508299

Pedder, D. (2006). Are small classes better? Understanding relationship between class size, classroom processes and pupils' learning. *Oxford Review of Education, 32*(2)*,* 213–234. https://doi.org/10.1080/03054980600645396

Peetsma, T., van der Veen, I., Koopman, P., & van Schooten, E. (2006). Class Composition Influences on Pupils' Cognitive Development. *School Effectiveness and School Improvement, 17*(3)*,* 275–302. https://doi.org/10.1080/13803610500480114

Peetsma, T., Vergeer, M., Roeleveld, J., & Karsten, S. (2001). Inclusion in Education: comparing pupils'development in special and regular education. *Educational Review, 53*(2), 125–135. https://doi.org/10.1080/00131910125044

Piaget, J., & Inhelder, B. (1956). *The Child's Conception of Space.* (Translated from French by F. J. Langdon and J. L. Lunzer). London: Routledge & Kegan Paul.

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H., & Weinert, S. (2016). Testing students with special educational needs in large-scale assessments - psychometric properties of test scores and associations with test taking behavior. *Frontiers in Psychology, 7*(154), 1–14. https://doi.org/10.3389/fpsyg.2016.00154

Postel, H. (1937). The Special School versus the Special Class. *Exceptional Children, 4*(1), 12–19. https://doi.org/10.1177/001440293700400103

Pulkkinen, J., & Jahnukainen, M. (2016). Finnish Reform of the Funding and Provision of Special Education: The Views of Principals and Municipal Education Administration. *Educational Review, 68*(2), 171–188. https://doi.org/10.1080/00131911.2015.1060586

Rea, P. J., McLaughlin, V. L., & Walter-Thomas, C. (2002). Outcomes for Students with Learning Disabilities in Inclusive and Pullout Programs. *Exceptional Children, 68*(2), 203–223.

Reynolds, D., De Fraine, B., Sammons, P., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): a state-of-the-art review School Effectiveness and School Improvement: *An International Journal of Research, Policy and Practice, 25*(2), 197–230. https://doi.org/10.1080/09243453.2014.885450

Rice, J. K. (1999). The impact of class size on instructional strategies and the use of time in high school mathematics and science courses. *Educational Evaluation and Policy Analysis, 21*(2), 215–229. https://doi.org/10.3102/01623737021002215

Richardson, J. & Powell, J. (2011). *Comparing special education. Origins to contemporary paradoxes*. Stanford, CA: Stanford University Press.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rojewski, J. W., Pollard, R. R., & Meers, G. D. (1990). Practice and attitudes of secondary industrial education teachers toward students with special needs. *Journal of industrial teacher education, 27*(3), 17–32.

Rosenbaum, P. R., & Rubin. D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Ross, J. D., & Ross, C. M. (1979). *Ross test of Higher Cognitive Processes.* Novato, CA: Academic Therapy Publications.

Ruijs, N. (2017). The impact of special needs students on classmate performance. *Economics of Education Review, 58,* 15–31. https://doi.org/10.1016/j.econedurev.2017.03.002

Ruijs, N., & Peetsma, T. (2009). Effects of inclusion on students with and without special educational needs reviewed. *Educational Research Review, 4*, 67–79. https://doi.org/10.1016/j.edurev.2009.02.002

Ruijs, N., Peetsma, T., & van der Veen, I. (2010.) The presence of several students with special educational needs in inclusive education and the functioning of students with special educational needs, *Educational Review, 62,* 1–37. https://doi.org/10.1080/00131910903469551

Ruijs, N., van der Veen, I., & Peetsma, T. (2010). Inclusive education and students without special educational needs. *Educational Research, 52,* 351–390. https://doi.org/10.1080/00131881.2010.524749

Schanzenbach, D. W. (2010). The Economics of Class size. *International Encyclopedia of Education (Third Edition),* 443–449. https://doi.org/10.1016/B978-0-08-044894-7.01236-7

Scharenberg, K. (2016). The Interplay of Social and Ethnic Classroom Composition, Tracking, and Gender on Students' School Satisfaction.

Journal of Cognitive Education and Psychology, 15(2), 320–346. https://doi.org/10.1891/1945-8959.15.2.320

Scharenberg, K., Rollett, W., & Bos, W. (2019). Do differences in classroom composition provide unequal opportunities for academic learning and social participation of SEN students in inclusive classes in primary school? *School Effectiveness and School Improvement, 30*(3), 309–327. https://doi.org/10.1080/09243453.2019.1590423

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment, 26*(4), 284–292. https://doi.org/10.1027/1015-5759/a000038

Schwab, S., & Hessels, M. (2015). Achievement Goals, School Achievement, Self-Estimations of School Achievement, and Calibration in Students With and Without Special Education Needs in Inclusive Education. *Scandinavian Journal of Educational Research, 59*, 461–477. https://doi.org/10.1080/00313831.2014.932304

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Company.

Shayer, M. (1979). Has Piaget's construct of formal operational thinking any utility? *British Journal of Educational Psychology, 49*, 265–276. https://doi.org/10.1111/j.2044-8279.1979.tb02425.x

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75,* 457–490. https://doi.org/10.3102/00346543075004457

Slavin, R. E. (1989). Class size and student achievement: Small effects of small classes. *Educational Psychologist, 24*(1), 99–110. https://doi.org/10.1207/s15326985ep2401_4

Slavin, R. E. (1990). Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis. *Review of Educational Research, 60*(3), 471–499. https://doi.org/10.3102/00346543060003471

Smith E., & Douglas, G. (2014). Special educational needs, disability and school accountability: an international perspective. *International Journal of Inclusive Education, 18*(5), 443–458. https://doi.org/10.1080/13603116.2013.788222

Snow, R. (1989). Aptitude, instruction, and individual development. *International Journal of Educational Research, 13*(8), 869–881. https://doi.org/10.1016/0883-0355(89)90070-0

Sternberg, R., Castejon, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. (2001). Confirmatory factor analysis of the Sternberg Triarchic Abilities Test in three international samples. *European Journal of Psychological Assessment, 17*, 1–16. https://doi.org/10.1027//1015-5759.17.1.1

Stringher, C. (2014). What is learning to learn? A learning process and output model. In R. Crick, C. Stringer, & K. Ren (Eds.), *Learning to Learn: International Perspectives from Theory and Practice* (pp. 1–33). London: Routledge.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science: *A Review Journal of the Institute of Mathematical Statistics, 25*(1), 1–21. https://doi.org/10.1214/09-STS313.

Sundqvist, C. Björk-Åman, A., & Ström, K. (2019). The three-tiered support system and the special education teachers' role in Swedish-speaking schools in Finland. *European Journal of Special Needs Education, 34*(5), 601–616. https://doi.org/10.1080/08856257.2019.1572094

Szumski, G., Smogorzewska, J., & Karwowski, M. (2017). Academic achievement of students without special educational needs in inclusive classrooms: A meta-analysis. *Educational Research Review, 21,* 33–54. https://doi.org/10.1016/j.edurev.2017.02.004

Tapola, A., & Niemivirta, M. (2008). The role of achievement goal orientations in students' perceptions of and preferences for classroom environment. *British Journal of Educational Psychology, 78,* 291–312. https://doi.org/10.1348/000709907X205272

Tarka, P. (2017). An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity, 52,* 313–354. https://doi.org/10.1007/s11135-017-0469-8

Teaching Qualifications Decree 986/1998. Amendments up to 105/2012. Government of Finland. Retrieved [15.10.2019] from https://www.finlex.fi/fi/laki/alkup/2012/20120105

Televantou, J., Marsha, H. W., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L.-E. (2015). Phantom effects in school composition research: consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement 26*(1), 75–101. http://dx.doi.org/10.1080/09243453.2013.871302

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90–118. https://doi.org/10.1080/00273171.2011.540475.

Thrupp, M. (1995). The school mix effect: The history of an enduring problem in educational research, policy and practice. *British Journal of Sociology of Education, 16,* 183–204. https://doi.org/10.1080/0142569950160204

Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research, 37,* 483–504. https://doi.org/10.1016/S0883-0355(03)00016-8

Thuneberg, H. (2007). *Is a majority enough? Psychological well-being and its relation to academic and prosocial motivation, self-regulation and achievement at school.* Research Report 281. Helsinki: University of Helsinki.

Thuneberg, H., Hautamäki, J., & Hotulainen, R. (2015). Scientific Reasoning, School Achievement and Gender: A Multilevel Study of between and within School Effects in Finland. *Scandinavian Journal of Educational Research, 59*(3), 337–356. https://doi.org/10.1080/00313831.2014.904426

Thuneberg, H., Vainikainen, M.-P., Ahtiainen, R., Lintuvuori, M., Salo, K., & Hautamäki, J. (2013). Education is special for all: The Finnish support model. *Gemeinsam leben, 2,* 67–78.

Trade Union of Education. (2019). *Oppimisen ja koulunkäynnin tuki uudistettava pikaisesti.* A press release. Retrieved [1.10.2019] from https://www.oaj.fi/ajankohtaista/uutiset-ja-tiedotteet/2019/oppimisen-ja-koulunkaynnin-tuki-uudistettava-ikaisesti/?fbclid=IwAR0wBDpw_eTyyPraDxuoUvCxFY6YerL1QOlubFW0FTWXvM64ixU2g6Z35qo

Turner, H. M., & Bernard, R. M. (2006). Calculating and Synthesizing Effect Sizes. *Contemporary issues in communication science and disorders, 33,* 42–55. https://doi.org/10.1044/cicsd_33_S_42

Törmänen, M., & Roebers, C. (2017). Developmental outcomes of children in classes for special educational needs: Results from a longitudinal study. *Journal of Research in Special Educational Needs, 18*(2), 83–93. https://doi.org/10.1111/1471-3802.12395

UNESCO. (1994). *The Salamanca Statement and Framework for Action on Special Needs Education*. Retrieved [20.8.2019] from: https://unesdoc.unesco.org/ark:/48223/pf0000098427

UNESCO. (2012). *International Standard Classification of Education.* ISCED 2011. Quebec: UNESCO.

UNESCO. (2017). *A guide for ensuring inclusion and equity in education*. Paris: UNESCO.

United Nations. (2006). *Convention on the Rights of Persons with Disabilities.* A/61/611. Retrieved [21.8.2019] from: http://www.un.org/esa/socdev/enable/rights/convtexte.htm#convtext.

Vainikainen, M.-P. (2014). *Finnish primary school pupils' performance in learning to learn assessments: A longitudinal perspective on the educational equity*. Research Report 360. Helsinki: University of Helsinki.

Vainikainen, M-P., & Hautamäki, J. (in press). Three Studies on Learning to Learn in Finland: Anti-Flynn Effects 2001.

Vainikainen, M-P., & Hautamäki, J. (2018). Selittääkö yrittäminen oppilaiden osaamisessa havaittuja ryhmäeroja? Itsearvioitu yrittäminen, investoitu työaika ja osaamiserot lokitietoanalyysin valossa. *Psykologia, 53*(2-3), 152–165.

Vainikainen, M.-P., Hautamäki, J., Hotulainen, R., & Kupiainen, S. (2015). General and specific thinking skills and schooling: Preparing the mind to new learning. *Thinking Skills and Creativity, 18*, 53–64. https://doi.org/10.1016/j.tsc.2015.04.006

Vainikainen M.-P., & Rimpelä, A. (2015). *Nuorten kehitysympäristö muutoksissa. Peruskoulujen oppimistulokset ja oppilaiden hyvinvointi eriytyvällä Helsingin seudulla.* Tutkimuksia 363. Helsinki: Helsingin yliopisto.

Vainikainen, M.-P., Thuneberg, H., Marjanen, J., Hautamäki, J., Kupiainen, S., & Hotulainen, R. (2017). How do Finns know? Educational monitoring without inspection and standard-setting. In S. Blömeke, & J.-E.

Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective* (pp. 243–259). (Methodology of educational measurement and assessment). Cham: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-50856-6_14

Van de gaer, E., Pustjens, H., Van Damme, J., & De Munter, A. (2006). The Gender Gap in Language Achievement: The Role of School-Related Attitudes of Class Groups. *Sex Roles, 55,* 397–408. https://doi.org/10.1007/s11199-006-9092-1

van Hek, M., Kraaykamp,G., & Pelzer, B. (2017). Do schools affect girls' and boys' reading performance differently? A multilevel study on the gendered effects of school resources and school practices. *School Effectiveness and School Improvement, 29,* 1–21. https://doi.org/10.1080/09243453.2017.1382540

Varjo J., & Kalalahti, M. (2019). The art of governing local education markets—municipalities and school choice in Finland. *Education Inquiry, 10*(2), 151–165. https://doi.org/10.1080/20004508.2018.1514907

Vettenranta, J., Välijärvi, J., Ahonen, A., Hautamäki, J., Hiltunen, J., Leino, K., Lähteinen, S., Nissinen, K., Nissinen, V., Puhakka, E., Rautopuro, J., & Vainikainen, M.-P. (2016). *Huipulla pudotuksesta huolimatta: PISA 15 ensituloksia.* Opetus- ja kulttuuriministeriön julkaisuja 2016:41. Helsinki: Opetus ja kulttuuriministeriö.

Virtanen, T., Haverinen, K., & Leskinen, M. (2018). Rakenneyhtälömallinnuksen menetelmällisiä ja käsitteellisteoreettisia lähtökohtia käyttäytymistieteellisessä tutkimuksessa. *Psykologia, 53*(4), 262–284.

Vygotsky, L. S. (1978). In M. Cole (Ed.), *Mind in society: The development of higher mental processes.* Cambridge, MA: Harvard University Press.

Wang, M. C., & Baker, E. T. (1985). Mainstreaming Programs: Design Features and Effects. *Journal of Special Education, 19*(4), 503–521. https://doi.org/10.1177/002246698501900412

Wechsler, D. (1981). *WAIS-R: Manual: Wechsler Adult Intelligence Scale–revised.* Harcourt Brace Jovanovich for Psychological Corp.

Wilkinson, I., Hattie, J., Parr, J., Townsend, M., Fung, I., Ussher, C., Thrupp, M., Lauder, H., & Robinson, T. (2000). *Influence of Peer Effects on Learning Outcomes: A Review of the Literature. Final report to the Ministry of Education*. Auckland, New Zealand: University of Auckland Uniservices. Retrieved [26.7.2019] from: http://files.eric.ed.gov/fulltext/ED478708.pdf

Wilkinson, I., Parr, J., Fung, I., Hattie, J., & Townsend, M. (2002). Discussion: modeling and maximizing peer effects in school. *International Journal of Educational Research, 37*, 521–535. https://doi.org/10.1016/S0883-0355(03)00018-1

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the raven progressive matrices test. *Computers in Human Behavior, 22,* 791–800. https://doi.org/10.1016/j.chb.2004.03.005

World Bank. (2019). *Inclusive Education Initiative: Transforming Education for Children with Disabilities.* Retrieved [26.7.2019] from www.worldbank.org/inclusive-education-initiative.

Wouters, S., Colpin, H., Van Damme, J., & Verschueren, K. (2013). Endorsing achievement goals exacerbates the big-fish-little-pond effect on academic self-concept. *Educational Psychology, 35*(2), 252–270, https://doi.org/10.1080/01443410.2013.822963

Yang Hansen, K., Gustafsson, J., & Rosen, M. (2014). *School Performance Differences and Policy Variations in Finland, Norway and Sweden. In Northern Lights on TIMSS and PIRLS 2011*. TemaNord 2014:528. Nordic Council of Ministers.

Yeah, S. (2009). Class size reduction or rapid formative assessment? A comparison of cost-effectiveness. *Educational Research Review 4*(1), 7–15. https://doi.org/10.1016/j.edurev.2008.09.001

YLE. (18.2.2019). *Suomi siirsi erityisoppilaat suuriin luokkiin, eivätkä kaikki opettajat pidä muutoksesta: "En ole koskaan ollut näin väsynyt"*. Retrieved [3.1.2020] from: https://yle.fi/uutiset/3-10644741

Zarghami Z., & Schnellert, G. (2004). Class Size Reduction: No Silver Bullet for Special Education Students' Achievement. *International Journal of Special Education, 19*(1), 89–96.

Zigmond, N. P., & Kloo, A. (2017). General and Special Education Are (and Should Be) Different. In J. M. Kauffman, D. P. Hallahan, & P. Cullen Pullen (Eds.), *Handbook of Special Education* (pp. 249–261). New York, NY: Education Routledge.

Zimmer, R., & Toma, E. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management, 19*, 75–92. https://doi.org/10.1002/(SICI)1520-6688(200024)19:1<75::AID-PAM5>3.0.CO;2-W

Zweers, I., Tick, N. T., Bijstra, J. O., & van de Schoot, R. (2019). How do included and excluded students with SEBD function socially and academically after 1,5 year of special education services? *European Journal of Developmental Psychology*, *17*(3), 317–335. https://doi.org/10.1080/17405629.2019.1590193